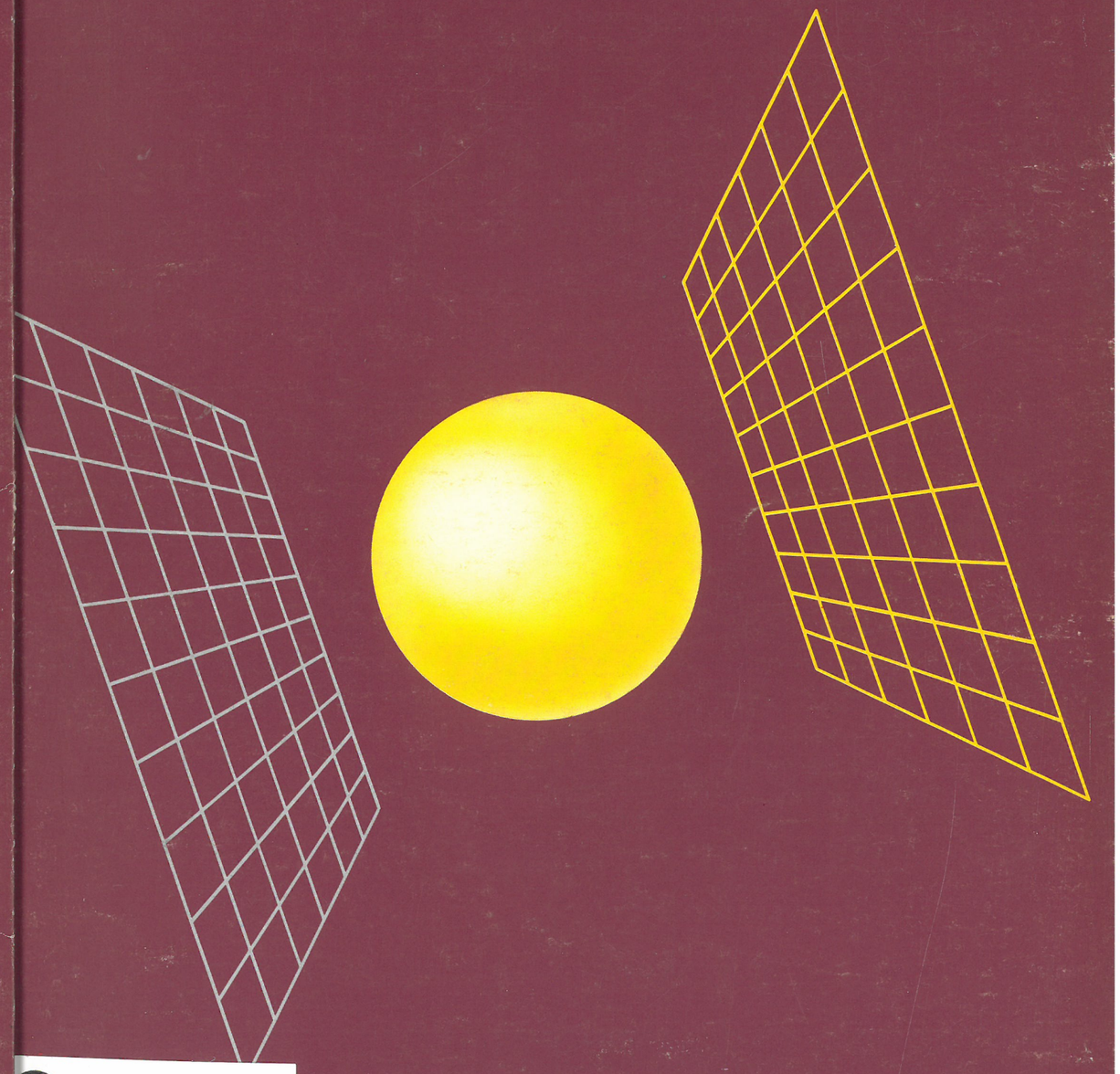
Spedizione in abbonamento postale a tariffa intera/Milano

Quaderni di informatica 28

assegna di tecnologia ed applicazioni degli elaboratori elettronici – Anno XII – Numero 3 – 1986



CENTRO STUDI INFORMATICA E AUTOMAZIONE
Honeywell
Honeywell Information Systems Italia

FPDM-35

Honeywell

Honeywell Information Systems Italia

Quaderni di informatica

Rassegna di tecnologia ed applicazioni degli elaboratori elettronici

Anno XII - Numero 3 - 1986

Sommario

Redazione
Honeywell Information Systems Italia
Centro Studi Informatica e Automazione
Via G.M. Vida, 11 - 20127 Milano

Stampa
tipolito Maggioni s.a.s.

Autorizzazione del Tribunale di Milano
n. 93 del 20 Marzo 1974

Oltre la macchina di Von Neumann

FRANCO FILIPPAZZI
GIULIO OCCHINI

Centro Studi Informatica e Automazione
Honeywell Information Systems Italia

1 - Introduzione

Col termine architettura ("computer architecture") si intende la struttura funzionale di un sistema di elaborazione, ossia i moduli (hardware e software) predisposti per l'esecuzione logica dei programmi.

Il termine non riguarda quindi gli aspetti tecnologici della realizzazione, ma ne designa lo schema funzionale, incluso una serie di elementi che lo caratterizzano, quali la lunghezza della parola, il numero di istruzioni di macchina, i modi di accesso alla memoria, etc.

Naturalmente, l'architettura non è indipendente dalla tecnologia, ma c'è tra loro lo stesso rapporto che esiste, ad esempio, tra il progetto di un edificio ed i materiali da costruzione.

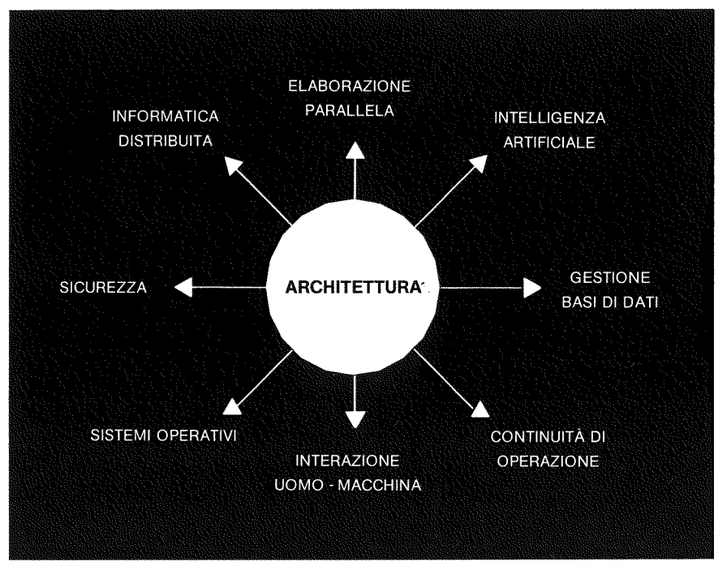
L'architettura di un elaboratore viene ovviamente progettata in base agli obiettivi funzionali e di costo che si vogliono ottenere. Alcuni requisiti possono condizionare o caratterizzare l'intero progetto architettonico; così esso può essere orientato ad ottenere grandissime capacità di elaborazione (super-computer), oppure sistemi che funzionino anche in presenza di guasto, o computer specializzati a gestire basi di dati, e così via. La figura 1 fornisce, a titolo indicativo, un ventaglio di direzioni lungo cui si sviluppa la ricerca architettonica; in questo articolo si farà una sintesi dei concetti relativi ai grandi computer ed alla elaborazione parallela.

2 - Tassonomia dei sistemi di elaborazione

Esaminando i sistemi di elaborazione esistenti o in sviluppo ci si trova di fronte ad una molteplicità di soluzioni diverse, in apparenza difficilmente inquadrabili in una classificazione organica, che permetta di fare confronti ed effettuare analisi.

Vari studi sono stati fatti per trovare un sistema di classificazione ed uno particolarmente significativo è quel-

Fig. 1 - Principali direzioni di sviluppo della architettura dei sistemi di elaborazione



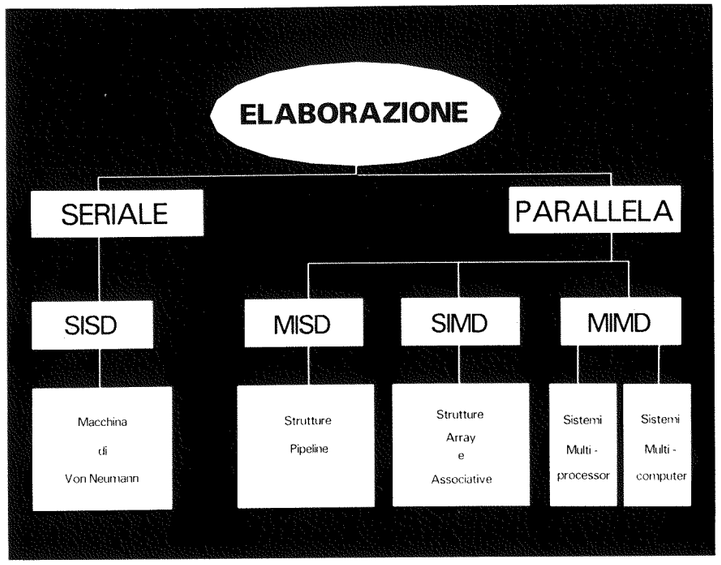
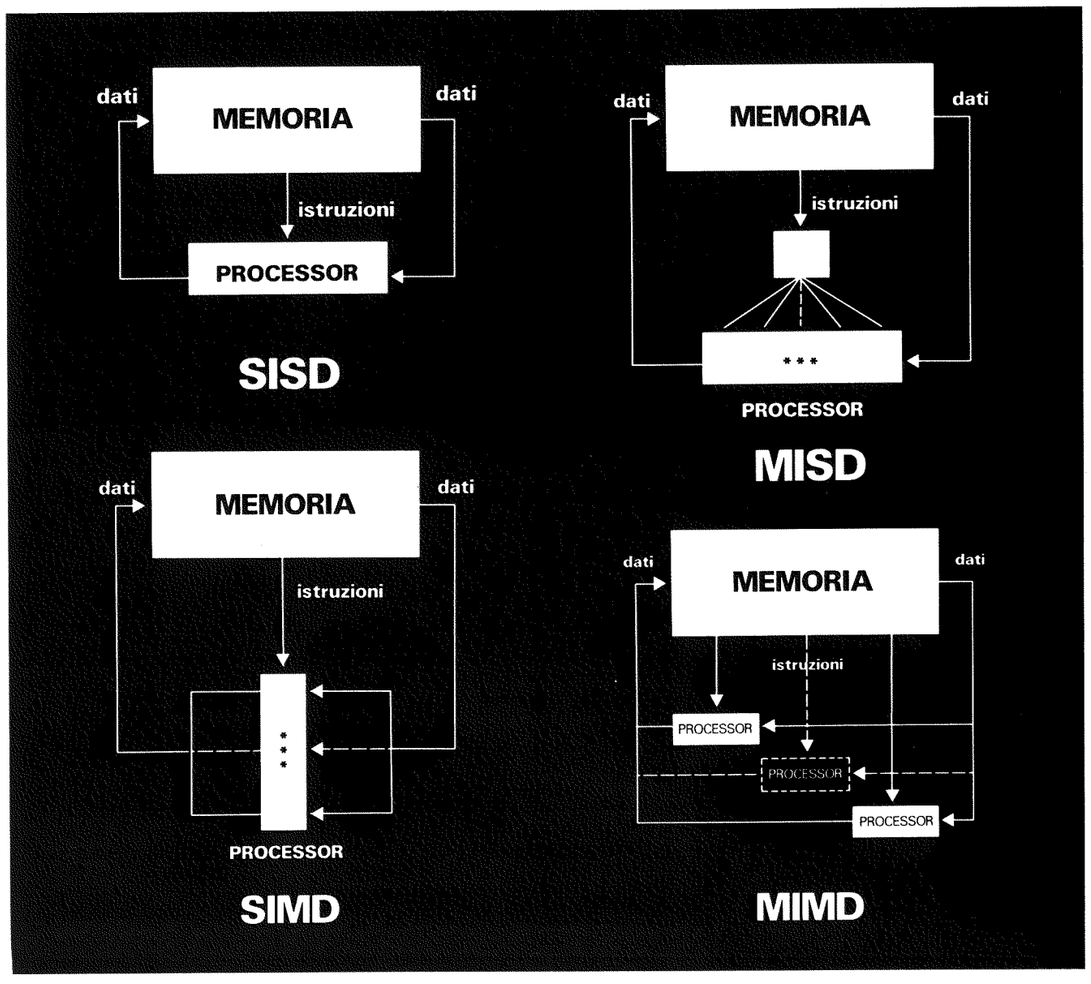


Fig. 2 - Architettura dei sistemi di elaborazione: tassonomia generale

Fig. 3 - Tassonomia delle architetture dei sistemi di elaborazione.



lo proposto da Flynn e sintetizzato nelle figure 2 e 3.

Gli elementi di base presi in considerazione in questa tassonomia sono (1) il numero di istruzioni eseguite simultaneamente dal computer e (2) il numero di operandi su cui esso opera in parallelo.

In effetti, in un sistema di elaborazione si possono individuare due fondamentali flussi di informazioni: uno di istruzioni ("instruction stream") che fluisce dalla memoria principale verso l'unità di elaborazione (processor), e l'altro di dati ("data stream") che fluisce tra la memoria e l'unità di elaborazione.

Da questo punto di vista, i computer possono essere classificati in quattro grandi classi a seconda che i flussi suddetti siano costituiti da una singola informazione o da una molteplicità di esse. Più precisamente, si possono distinguere le seguenti categorie di macchine:

- SISD (Single Instruction Stream, Single Data Stream)
- MISD (Multiple Instruction Stream, Single Data Stream)
- SIMD (Single Instruction Stream, Multiple Data Stream)
- MIMD (Multiple Instruction Stream, Multiple Data Stream)

Le architetture SISD costituiscono la soluzione classica, che caratterizza tuttora la grande maggioranza delle macchine. Il sistema ha un unico processor (cioè una unità di controllo ed una unità logico-aritmetica, o ULA) e le istruzioni vengono eseguite serialmente, estraendo un dato per volta dalla memoria. Questo schema sta all'origine dei computer ed è perciò detto anche di von Neumann.
Nelle architetture MISD il processor è costituito, oltre che dalla unità di controllo, da una serie di ULA, disposte in serie, ognuna delle quali esegue un'istruzione diversa su uno stesso flusso di dati. Questa struttura è nota anche come "pipeline", perché le ULA costituiscono i segmenti di un "condotto", attraverso cui fluiscono i dati.
Il vantaggio di questa struttura deriva dal fatto che non si aspetta che un dato sia completamente trattato prima di passare al successivo.
Un'analogia intuitiva è quella di una catena di montaggio; come in questa, però, affinché il flusso sia continuo, occorre cambiare il meno possibile le operazioni effettuate in ogni singola stazione. La struttura pipeline è perciò particolarmente utile nel caso di programmi con elevata ripetitività, che eseguono cioè la stessa sequenza di operazioni un rilevante numero di volte.

Nelle architetture SIMD il processor è costituito ancora da una molteplicità di ULA, le quali però eseguono la medesima istruzione contemporaneamente su operandi diversi. Queste unità sono spesso disposte fisicamente in una configurazione a matrice ("array processor"), con collegamenti diretti tra ciascuna di esse e i suoi immediati vicini.

Schemi di questo genere sono tipici delle grosse macchine scientifiche attuali. Essi risultano efficienti in problemi i cui dati presentano strutture regolari (manipolazione di matrici, elaborazione d'immagini, simulazione di fenomeni fisici, etc.), in cui la configurazione della matrice di ULA è, in un certo senso, speculare del problema. Queste macchine sono spesso usate in unione con un computer di tipo tradizionale che ne controlla il funzionamento complessivo.

L'ultima classe di macchine, la MIMD, è quella che presenta le maggiori potenzialità e versatilità.

Questi sistemi sono infatti costituiti da un insieme di unità general-purpose asincrone, ciascuna in grado di operare indipendentemente dalla altre, ma anche di cooperare con esse alla soluzione di uno stesso problema.

Si distinguono, in effetti, due sottoclassi di macchine MIMD, e cioè sistemi "ad accoppiamento stretto", in cui i processor accedono direttamente alle risorse del sistema (in particolare alle memorie); e sistemi "ad accoppiamento lasco" in cui tale accesso avviene secondo meccanismi formalizzati. Nel primo caso si parla di sistema "multiprocessor", che costituisce di solito la struttura interna di una singola macchina fisica; nel secondo caso si parla invece di sistema "multicomputer" (o "rete di computer"), che è costituito invece da una pluralità di macchine delle classi viste in precedenza, sparse geograficamente, collegate tra loro da mezzi vari di telecomunicazione. Nella classe MIMD si possono far rientrare alcune delle architetture più innovative attualmente in sviluppo, cui si accennerà nel seguito (macchine dataflow, ecc.).
Dopo questa introduzione di inquadramento sulle architetture, possiamo meglio esaminare gli orientamenti in corso.

3 - I supercomputer

Un orientamento fondamentale, su cui si investono molte risorse, è quello diretto a realizzare grandi capacità di elaborazione. La fig. 4 mostra, in proposito, come ad ogni generazione di computer sia associato un salto di prestazioni. Ma è realmente necessario costruire macchine ultrapotenti, i cosiddetti "supercomputer"? In effetti ci sono numerose applicazioni in cui la potenza di calcolo occorrente costituisce attualmente una limitazione; a titolo di esempio, citiamo due casi in settori diversi. L'elaborazione dei fotogrammi (cioè i complessi procedimenti matematici per migliorare la qualità dell'immagine eliminando il "rumore", correggendo le distorsioni introdotte dai sistemi ottici, etc.) richiede calcoli laboriosissimi: ad esempio, un singolo fotogramma del tipo di quelli inviati dai satelliti per l'esplorazione dello spazio richiede molte ore di lavoro del più potente elaboratore oggi disponibile. Se si pensa alla quantità di immagini che vengono inviate quotidianamente dallo spazio sulla Terra, ci si rende conto dell'interesse a disporre di supercomputer. Un altro esempio è rappresentato dalla previsione meteorologica a lungo termine. Usando il modello più completo, che descrive la circolazione atmosferica del globo tenendo conto dell'influenza dei mari e delle nubi, si richiede, per la previsione sull'arco di qualche mese, un volume di calcoli che è assolutamente al di fuori di ogni possibilità attuale. Molte altre applicazioni potrebbero essere citate, in campo scientifico, tecnico, militare, che richiedono elaboratori

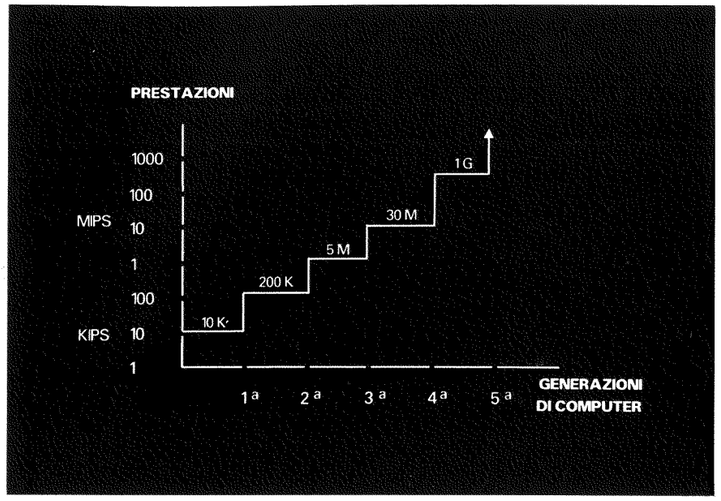


Fig. 4 - L'aumento della potenza di elaborazione.

centinaia o migliaia di volte più potenti di quelli attuali. La stessa possibilità di evoluzione del computer verso macchine "intelligenti" è condizionata (oltre che da altri fattori) dalla disponibilità di grandi potenze di elaborazione.

4 - L'elaborazione parallela

Come viene affrontato tecnicamente il problema dei grandi elaboratori? In sostanza, ci si muove contemporaneamente su due fronti: nell'area tecnologica, per realizzare circuiti sempre più veloci e nell'area sistemistica, per realizzare architetture largamente parallele.

Le nuove architetture su cui si lavora differiscono profondamente dalla macchina di von Neumann, tanto che un modo corrente per indicarle in blocco è appunto quello di "macchine non -von Neumann".

Nell'architettura di von Neumann esiste un collo di bottiglia fondamentale, rappresentato dalla comunicazione tra processor e memoria (fig. 5a). Essendo essi unici, possono scambiarsi solo una istruzione o un dato per volta. L'operazione è quindi intrinsecamente sequenziale; essa è inoltre "guidata dal programma", nel senso che l'unità di controllo legge ed esegue via via le istruzioni, nell'ordine in cui esse compaiono nel programma.

Il modo più generale di ovviare alla "strozzatura di von Neumann" è rappresentato in fig. 5b; si tratta di uno schema MIMD, in cui tanti processor identici vengono connessi contemporaneamente e dinamicamente tra loro e con le unità di memoria, attraverso una rete di commutatori elettronici (Architettura multiprocessor "shared memory").

Una variante di questo schema è quella di fig. 5c (multiprocessor "message passing"), in cui ad ogni processor è associata una memoria locale.

La realizzazione, cui attualmente si mira, di architetture parallele costituite da migliaia di processor e nelle quali avviene contemporaneamente una molteplicità di processi "concorrenti", rappresenta una sfida sul piano concettuale e realizzativo.
Esistono numerose variabili di progetto, quali il modo di:

- governare il sistema ("control")
- dimensionare i processor ("granularity")
- suddividere un programma ai fini della massima efficienza ("partitioning")
- assegnare le varie parti di un programma ai diversi processor ("scheduling")
- assicurare una sequenza di esecuzione che conduca a risultati corretti ("synchronizing")
- ecc.

Possiamo qui far cenno solo al primo problema, che risulta peraltro un aspetto chiave dell'intera architettura.

Il controllo del sistema può essere centralizzato o decentrato. Nel primo caso si va incontro in particolare a problemi di velocità associati con l'esistenza di lunghi percorsi di interconnessione; nel secondo caso si hanno invece problemi nel coordinare le risorse del sistema.

Esistono attualmente tre fondamentali approcci al controllo, indicati nella fig. 6; di essi uno ("control-driven") è di tipo centralizzato, mentre gli altri due ("data-driven" e "demand-driven") sono decentrati.

I sistemi multiprocessor control-driven (ossia pilotati dal controllo) appartengono in effetti alla classe SIMD. Di norma esiste un processor centrale di tipo convenzionale, che invia contemporaneamente l'istruzione da eseguire a tutti i processor del sistema. Si stanno progettando macchine sperimentali di questo tipo con un numero elevatissimo di processori e granularità molto piccola.

Le architetture concettualmente più innovative sono però quelle del tipo data-driven e demand-driven. In queste strutture, ogni istruzione da eseguire per ottenere il risultato finale non è attivata dal programma, come nell'impostazione classica, ma è innescata rispettivamente dalla presenza o dalla richiesta di dati. Si

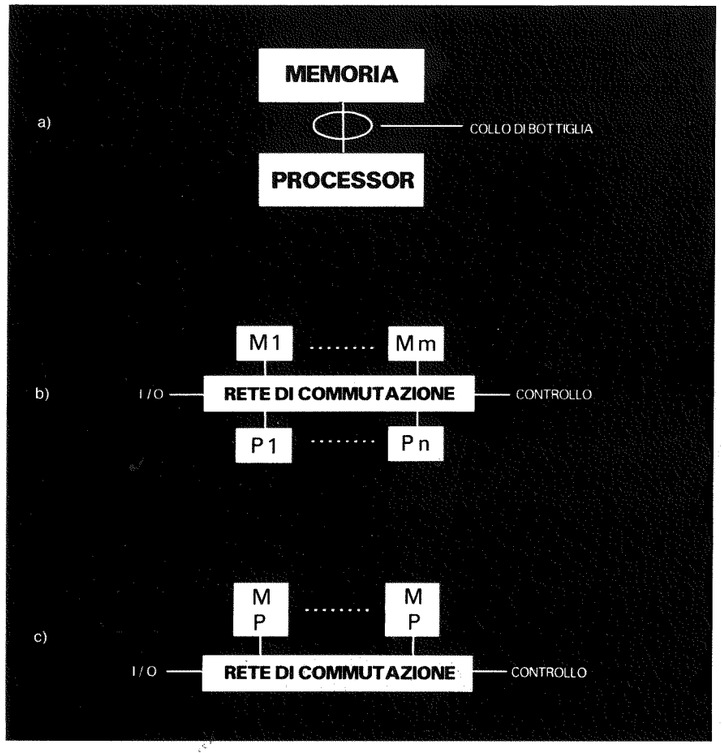


Fig. 5 - Oltre le macchine di von Neumann.
a) Macchina di von Neumann;
b) sistema multiprocessor shared-memory;
c) sistema multiprocessor message-passing.

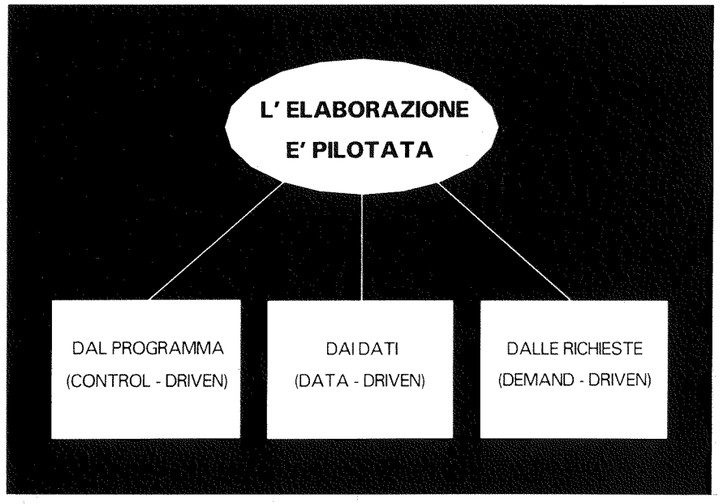


Fig. 6 - Architetture parallele: tipologia di controllo

può meglio chiarire il concetto con un esempio.
Si debba calcolare l'espressione (a+b) x (c-d). In una macchina di von Neumann il processo è completamente seriale; occorre infatti prima calcolare (a+b) e mettere il risultato in memoria; poi calcolare (c-d) e memorizzare il risultato; e infine richiamare i due risultati parziali e moltiplicarli tra loro.

In una macchina data-driven la presenza delle varibili a,b,c,d, innesca automaticamente l'operazione di somma (a+b) e di sottrazione (c-d), che vengono eseguite simultaneamente da due processor diversi; i risultati innescano a loro volta, appena pronti entrambi, un altro processor che esegue la moltiplicazione. È evidente come questo modo di procedere sia molto più rapido di quello tradizionale perché vengono eseguite in parallelo tutte le operazioni che l'algoritmo consente (si elimina, inoltre, il "parcheggio" in memoria dei risultati parziali).

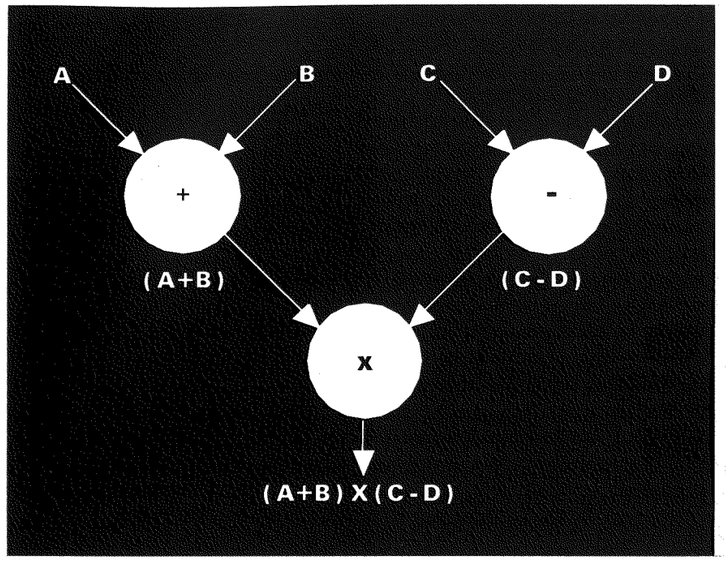


Fig. 7 - Rappresentazione di un programma mediante un grafo orientato. Le nuove architetture consentono di sfruttare il parallelismo intrinseco di un algoritmo.

Nelle macchine data-driven (dette anche macchine "data-flow") un programma può essere descritto mediante un grafo orientato, del tipo illustrato in fig. 7. Tale grafo viene, in sostanza materializzato dal sistema, interconnettendo dinamicamente le risorse hardware disponibili, man mano che lo sviluppo della elaborazione lo richiede. I dati percorrono gli archi del grafo, i cui nodi rappresentano funzioni da eseguire.

Quando in un nodo sono presenti tutti i dati richiesti, viene automaticamente eseguita la funzione corrispondente. Gli operandi spariscono allora dagli archi di ingresso, mentre un nuovo dato compare sull'arco di uscita del nodo.

Nelle macchine demand-driven le varie operazioni sono invece attivate dalla richiesta di risultati. In sostanza, il grafo che descrive il programma è percorso in senso inverso del caso data-driven: è infatti la richiesta di un risultato che fa scattare la computazione relativa, che a sua volta fa scattare la valutazione dei suoi argomenti, e così via. Questa propagazione di richieste continua finchè non si incontrano dei dati effettivi, anziché dei sottoprogrammi; tali valori sono allora rimandati indietro ai nodi richiedenti e l'esecuzione del programma procede in direzione opposta alla precedente.

Nelle macchine demand-driven, i programmi possono ancora essere descritti come grafi, in cui le richieste si propagano lungo gli archi nella direzione delle frecce e i dati nella direzione opposta. Si può anche dire che la funzione relativa ad un nodo viene eseguita quando il risultato viene richiesto da altri nodi.

Le macchine demand-driven (dette anche macchine "reduction") sono considerate con particolare interesse nell'area della intelligenza artificiale, in cui si ha normalmente a che fare con problemi in cui la conoscenza o i dati sono incompleti.

In questo caso, infatti, la soluzione può essere trovata mediante manipolazioni sui simboli quali nelle dimostrazioni algebriche.

È evidente da quanto detto che le nuove architetture rappresentano un cambiamento radicale rispetto agli schemi di computazione tradizionali. Infatti:

- non è più il programma che guida l'esecuzione delle operazioni, ma sono i dati (o la loro richiesta) che determinano la sequenza con cui esse vengono eseguite;
- manca il concetto di controllo centrale del processo; le elaborazioni vengono eseguite automaticamente non appena ci sono dati e risorse (processor) disponibili;
- il sistema può espletare tutto il parallelismo intrinseco di un algoritmo (ove vi sia un sufficiente numero di processor).

Chiaramente, il parallelismo operativo di queste macchine non dipende solo dal numero di processor che le costituiscono. A prescindere da altre ragioni di inefficienza, il numero di operazioni simultanee non può essere maggiore di quelle consentite dallo specifico algoritmo con cui si è risolto il problema. In altri termini, come ovvio, la velocità di esecuzione è legata anche alla ricerca di algoritmi paralleli.

Architetture non-von Neumann cominciano ad esistere a livello di macchine sperimentali con un numero limitato di processor. Si può prevedere negli anni a venire la realizzazione di macchine parallele con migliaia di processor. Un ruolo importante giocheranno, a tale fine, i progressi della microelettronica e la capacità di integrare molti processor nello stesso substrato. Ma occorrerà risolvere anche molti altri problemi che riguardano la struttura architettonica delle macchine e la logica della computazione.

5 - Strutture ipercubiche, sistoliche e associative

Il tema dei supercomputer e della elaborazione parallela merita ancora alcuni cenni.

Fig. 8 - Macchine parallele: alcune configurazioni di architetture sistoliche.

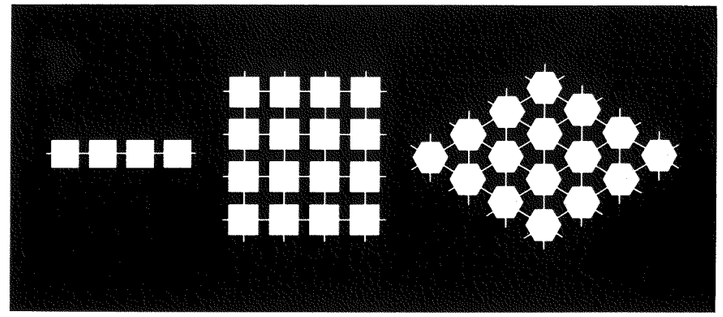
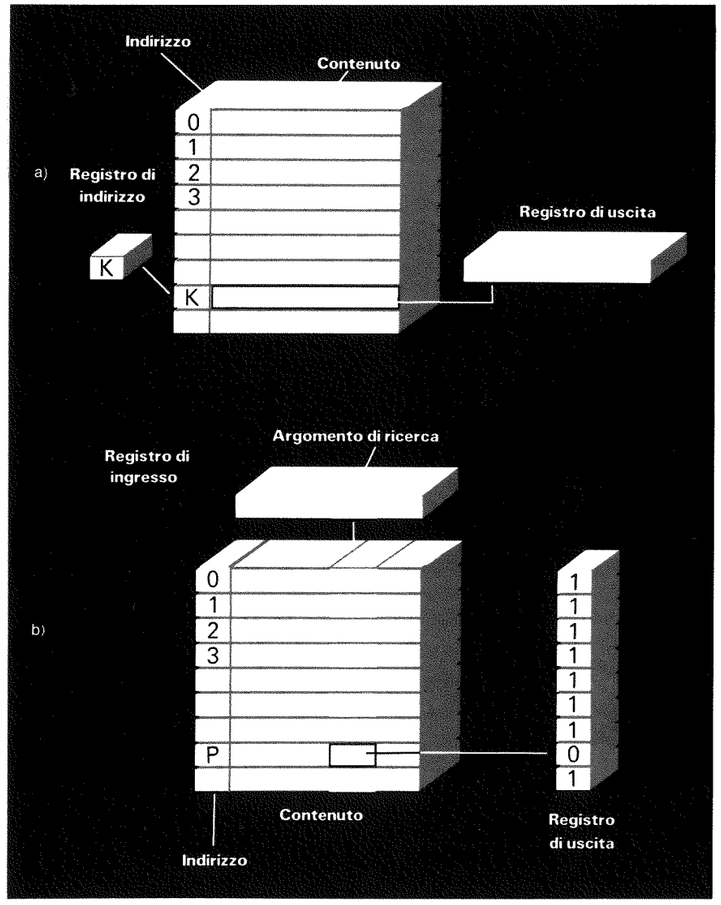


Fig. 9 - Memorie con funzionamento parallelo
a) memoria tradizionale, con accesso per posizione
b) memoria associativa, con accesso per contenuto



Un problema fondamentale nel progetto di macchine parallele è rappresentato dalla interconnessione dei processor. La soluzione più generale è, ovviamente, quella che consente a tutti i processor del sistema di comunicare direttamente tra loro. Questa soluzione è realizzata da strutture come quelle di fig. 5 b/c, viste in precedenza; in esse, una rete di commutatori elettronici svolge un ruolo paragonabile a quello di una centrale telefonica. Questa rete di commutazione costituisce un fattore critico del sistema, a causa della grande velocità di funzionamento richiesta, nonché per la sua complessità.

Una soluzione meno generale è rappresentata da strutture in cui ogni processor è connesso soltanto con quelli ad esso adiacenti. I processor possono essere collocati nei nodi di una griglia o nei vertici di un cubo o anche in strutture più complesse, indicate come "ipercubi". Se N sono le dimensioni dell'ipercubo, ciascun processor risulta connesso con N altri; si constata immediatamente che in tale struttura un messaggio richie-

de un massimo di N passi per andare da un processor all'altro e che l'intero cubo è costituito da 2^N processor. Esistono attualmente progetti basati su un ipercubo a 16 dimensioni.

Nel quadro delle strutture parallele ora viste, va citata una interessante categoria di macchine dette "sistoliche". Anche in questo caso, si ha un insieme di processor, ognuno piuttosto semplice e connesso con quelli circostanti secondo geometrie regolari (si vedano alcuni esempi in fig. 8). Caratteristica del funzionamento è che il flusso di dati che attraversa il reticolo di processor è costante; esso viene infatti "pompato" ritmicamente dalla memoria, come il cuore fa col sangue, da cui il nome dato a questa architettura. Il rendimento della macchina è evidentemente massimizzato se ciascun dato che, ad ogni pulsazione, passa dall'una all'altra cella della struttura, viene da esse effettivamente utilizzato. È intuitivo che esiste una stretta correlazione tra la specifica topologia della macchina e le classi di algoritmi che possono venire trattati in modo efficiente.

Il problema di usare efficientemente strutture parallele è, d'altra parte, di carattere generale. Scrivere i programmi in modo che sfruttino la capacità di un elevato insieme di processor è tutt'altro che semplice; occorre, in particolare, suddividere il problema in opportune sottoparti e non sempre ciò è fattibile in modo adeguato.

Sulla possibilità di operazione parallela c'è da aggiungere che il concetto si applica oltre che ai sistemi di elaborazione in senso lato, anche a quell'organo fondamentale di ogni sistema che è la memoria.

Tutte le memorie sino ad ora impiegate sono caratterizzate dal fatto che per accedere ad un dato se ne deve conoscere l'"indirizzo", ossia la posizione da esso occupata nella memoria.

È però possibile concepire un tipo di memoria completamente diverso, in cui per accedere ad un dato non è più necessario conoscerne l'indirizzo. Questo tipo di memoria viene chiamato "associativa" o anche "memoria indirizzata per contenuto" (CAM = Content Addressed Memory); v. fig. 9B.

Si può chiarire il concetto con un esempio. Si abbia una memoria contenente i dati anagrafici di una popolazione (ossia, per ogni individuo, il cognome, il nome, il luogo di nascita, ecc.) e si supponga di voler individuare tutte le persone che, per esempio, abbiano una data età.

Utilizzando una memoria tradizionale è necessario esplorare una per tutte le posizioni della memoria, confrontando ogni volta l'età della persona con il valore dato. Occorre cioè eseguire tante interrogazioni quante sono le persone.

Se invece si ha una memoria associativa basta impostare il campo di ricerca (l'anno di nascita in questo caso) e la risposta viene fornita con una sola interrogazione.

Come può essere realizzata una memoria di questo tipo? In sostanza, accoppiando in ciascuna cella la funzione di memoria con delle capacità logiche. Infatti, ogni bit di memoria associativa è formato da un elemento di memoria convenzionale e da un circuito logico di confronto. I campi di ricerca vengono applicati simultaneamente a tutte le celle della memoria. Se in una cella c'è discordanza tra il bit memorizzato ed il bit di confronto, essa genera un segnale che viene memorizzato in un apposito registro; esaminando questo registro, si possono immediatamente individuare le posizioni di memoria cercate.

Come si vede, si tratta di un processo di ricerca altamente parallelo, perché si esamina tutta la memoria in un solo passo. La ricerca di dati in una memoria associativa risulta quindi estremamente più veloce che in una memoria convenzionale; inoltre, le informazioni possono essere messe nella memoria in modo casuale, perché non è necessario conoscerne la collocazione per rintracciarle.

È evidente l'interesse che può avere una memoria di questo tipo, in cui si accede in parallelo a tutte le informazioni, specificando "quale" informazione si desidera, anziché "dove" essa è collocata. Basti pensare alla ricerca documentaria in grandi archivi bibliografici, giuridici, brevettuali, ecc.

Ma il concetto di memoria associativa può essere utilmente impiegato nel funzionamento generale di un elaboratore, per accelerare il reperimento dei dati che via via vengono richiesti durante l'elaborazione. Memorie di questo tipo risultano quindi particolarmente utili per elaboratori di elevate prestazioni e in campi quale l'intelligenza artificiale, in cui occorre effettuare in continuazione ricerche in grandi "basi di conoscenza".

Attualmente, però, queste memorie sono pochissimo usate; praticamente solo nei grandi elaboratori si trovano piccole memorie associative (realizzate con circuiti integrati), affiancate per usi particolari alle grandi memorie convenzionali. La ragione è che, come si può arguire dalla loro complessità, queste memorie hanno costo, ingombro, dissipazione assai maggiori di quelle convenzionali. La microelettronica è la chiave tecnologica che potrà consentire un impiego pratico delle memorie associative, realizzando con ciò un'ulteriore importante evoluzione dell'elaboratore.

Il materiale di questo articolo è tratto dal volume degli stessi Autori: «Le frontiere dell'informatica», Edizioni del Sole/24 Ore, di recente pubblicazione (ottobre 1986).

Una breve presentazione dell'opera è fornita nella rubrica «Biblioteca Novità» in questo numero della rivista.

Olografia col computer

STEVE ARNOLD

Honeywell
Physical Sciences Center
Minneapolis (USA)

1. Introduzione

Un ologramma può essere considerato come una rappresentazione fisica, "congelata" nello spazio e nel tempo, di un fenomeno dinamico come la luce.

Una fotografia registra e rappresenta le variazioni nella intensità della luce riflessa da un oggetto; un ologramma cattura invece nella loro interezza (holos = intero) le onde luminose riflesse, la configurazione microscopica di picchi e valli delle onde elettromagnetiche. La configurazione originale delle onde può essere ricostruita facendo passare un fascio di luce laser attraverso l'ologramma.

L'immagine formata in questo modo appare tridimensionale perchè i fronti d'onda che raggiungono l'occhio dell'osservatore sono praticamente gli stessi che sarebbero prodotti se l'oggetto rappresentato dall'immagine fosse realmente presente.

L'immagine così prodotta può essere fotografata, persino registrata in un altro ologramma.

Il metodo convenzionale di produzione di un ologramma è strettamente un metodo di registrazione: come nel caso della fotografia il metodo può rappresentare ciò che già esiste. Tuttavia può nascere l'esigenza di creare un ologramma "ex novo" basato su una configurazione d'onda definita teoricamente, in accordo con una funzione matematica prestabilita.

Un ologramma di questo tipo risulterebbe analogo non ad una fotografia, ma ad un dipinto o a un disegno.

In linea di principio esso potrebbe rappresentare ogni sistema di fronti d'onda che può essere immaginato.

Le difficoltà tecniche nel realizzare un tale ologramma direttamente sono considerevoli. Un'area approssimativamente di un centimetro quadrato deve essere riempita con una figura in cui i più piccoli dettagli sono dimensionalmente dello stesso ordine di grandezza della lunghezza d'onda della luce visibile, o in altre parole inferiori al micron.

E ancora più stringenti sono le tolleranze relative al posizionamento di questi dettagli.

Metodi pratici ed efficaci sono necessari sia per calcolare i dati che definiscono la figura dell'ologramma, sia per disegnare effettivamente l'ologramma.

Per il calcolo, abbiamo fatto uso, come ci si sarebbe potuto aspettare, del sussidio di calcolatori elettronici.

Per il compito di tracciamento e disegno dell'ologramma abbiamo adottato invece una tecnologia sviluppata per la produzione dei circuiti integrati.

Ologrammi generati con l'ausilio del calcolatore hanno applicazioni potenziali come componenti di sistemi ottici, sostituendo lenti convenzionali, specchi, griglie di diffrazione e simili.

Una applicazione che è già stata utilizzata in pratica con successo è la verifica di componenti ottici convenzionali.

Una prospettiva più lontana, ma non remota o impossibile, è l'uso di ologrammi, generati mediante calcolatore, per l'esecuzione di calcoli per via ottica, dove fasci di luce modulata si sostituiscono ai segnali elettronici attuali.

Noi abbiamo creato ologrammi che svolgono operazioni matematiche fondamentali con mezzi ottici.

Sebbene gli ologrammi in se abbiano svolto il loro compito in modo soddisfacente, altri elementi del sistema di calcolo richiedono un ulteriore affinamento.

2. L'olografia convenzionale

I principi dell'olografia possono essere illustrati pensando di eseguire il seguente esperimento. Supponiamo di avere una sorgente di luce monocromatica e coerente, dove i diversi raggi di luce hanno la stessa lunghezza d'onda o colore, e sono tutti in fase, picchi allineati con picchi e valli con valli. Un laser è una sorgente di luce di questo tipo.

Il fascio luminoso è riflesso da un oggetto e poi, senza essere focalizzato da una lente, incide su una lastra fotografica. Poichè non ci sono lenti di focalizzazione, ogni punto del film riceve una porzione della luce diffusa da ciascun punto illuminato dell'oggetto.

Perciò il film registra puramente la

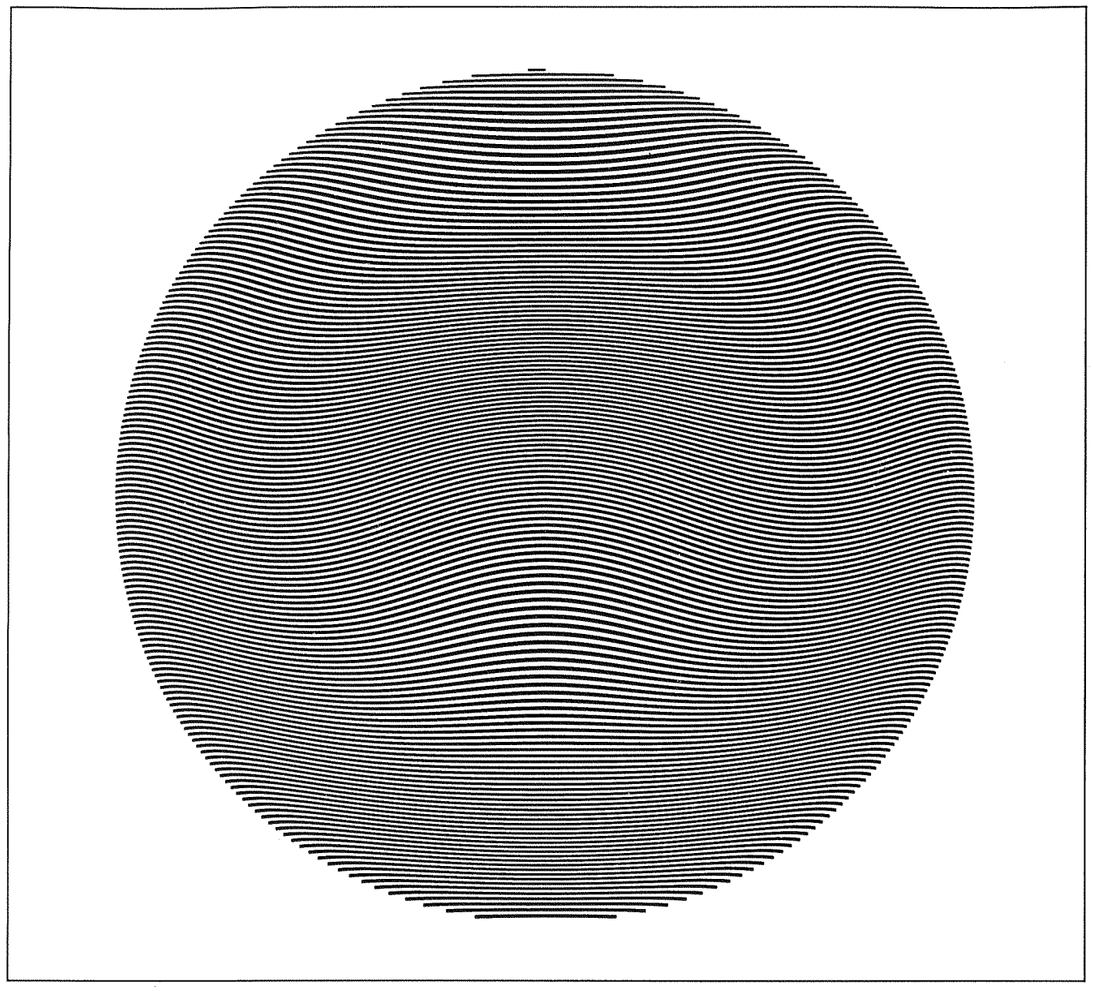


Fig. 1 - La figura rappresenta un ologramma generato col calcolatore, ingrandito fotograficamente. Gli ologrammi, che possono essere considerati come complicate griglie di diffrazione, possono sostituire i convenzionali elementi ottici rifrangenti in molte applicazioni. Un elaboratore viene usato per calcolare la figura di interferenza che genera un fronte d'onda desiderato. Questo ologramma riproduce il fronte d'onda generato da una lente asferica. L'ologramma è stato molto ingrandito per rendere possibile la stampa. Di fatto, un ologramma generato al calcolatore ha tipicamente la dimensione di una monetina e contiene molte più frange.

luminosità media della scena e, quando sviluppato, è di un grigio uniforme.

Risulta chiaro che nessuna informazione relativa all'oggetto illuminato è stata registrata nella fotografia.

Eppure, la luce che raggiunge il film contiene e convoglia tutte le informazioni relative all'oggetto, che potrebbero essere incorporate in una fotografia.

Supponiamo ora che le onde luminose possano essere in qualche modo "fermate" sul loro percorso.

Un demone microscopico potrebbe allora essere reclutato per registrare l'ampiezza e la fase di ogni onda luminosa nel punto in cui questa attraversa il piano del film.

Con queste informazioni l'immagine dell'oggetto potrebbe essere ricostruita in tutta la sua intierezza.

Tutto ciò che è necessario compiere per conseguire questo risultato è di far partire dal film un nuovo insieme di onde luminose, propagantesi con la stessa ampiezza iniziale e con la stessa fase che le onde riflesse avevano all'istante in cui erano state fermate nel film.

In pratica, ovviamente, le onde luminose non possono essere fermate nel loro percorso, ma un ologramma cattura le stesse informazioni in un altro modo. Il fascio di luce coerente è scomposto, alla sorgente, in due fasci e solo una delle due frazioni viene riflessa dall'oggetto.

L'altra viene fatta incidere direttamente sulla lastra fotografica.

Quando i due fasci di luce si incontrano sulla lastra, interferiscono tra di loro.

Dove due picchi o due valli si trovano in coincidenza, l'ampiezza dell'onda risultante viene aumentata; dove un picco e un avvallamento si trovano in coincidenza, le due onde si elidono reciprocamente.

Per di più, queste relazioni di fase non sono così fugaci da poter essere percepite solo in una immaginaria istantanea di onde congelate.

Poichè la luce è monocromatica e coerente, se il raggio riflesso e il raggio di riferimento sono in fase a un certo istante e ad un certo punto del film, essi in quel punto, resteranno indefinitamente in fase.

Ne consegue che il film risulterà impressionato nei punti in cui i due raggi interferenti sono in fase, non impressionato dove le onde incidenti sono in opposizione di fase.

La figura di interferenza creata in questo modo consiste in aree alternativamente chiare e scure che generalmente prendono la forma di finissime linee sinuose chiamate frange.

È questa figura che costituisce l'ologramma.

Quando una luce laser è fatta passare attraverso il film, la configurazione di linee finemente spaziate diffrange variamente le onde incidenti, cosicchè la luce trasmessa da ogni regione elementare del film si apre a ventaglio.

Al di là del piano dell'ologramma le onde diffratte interferiscono tra loro, riproducendo con dettaglio microscopico la configurazione di interferenze che aveva dato luogo all'ologramma.

Le interferenze attenuano le onde che si propagano in certe direzioni e ne amplificano altre.

L'effetto ultimo che ne risulta è una replica del fronte d'onda riflesso dall'oggetto e intercettato dal film.

3. Capacità di informazione

In un ologramma generato mediante calcolatore l'intero processo di registrazione viene aggirato ed evitato.

Il primo passo nel disegno dell'ologramma è la definizione del fronte d'onda che sarebbe generato dall'oggetto che si intende rappresentare. La specificazione prende forma matematica e può essere effettuata anche se l'oggetto stesso non esiste. Il passo successivo consiste nel calcolo del fronte d'onda in corrispondenza del piano dell'ologramma e nella sua conversione in una descrizione delle frange di interferenza, espressa in una forma adatta a comandare un dispositivo grafico di uscita. Il passo finale, ovviamente, consiste nel tracciamento effettivo dell'ologramma.

Un ologramma generato da calcolatore differisce per molti aspetti da un ologramma ottenuto per via ottica. Il processo ottico è per sua natura analogico, cosicchè l'intensità delle frange varia in modo continuo dallo scuro al chiaro.

La codifica digitale dei dati di interferenza richiede che a ciascun punto dell'ologramma sia attribuito un livello discreto di intensità o di grigio, e in pratica (per limitare il numero di dati da elaborare) noi abbiamo lavorato esclusivamente con ologrammi "binari" in cui ciascun punto è nero o bianco, opaco o trasparente. La figura dell'ologramma è costituita da punti elementari di immagine ("pixel") che possono assumere solo due valori.

Un indicatore chiave della qualità di un ologramma o delle sue prestazioni è il "prodotto spazio-ampiezza di banda". Esso si determina moltiplicando l'area totale dell'ologramma per un qualche misura della quantità di informazioni registrate sulla unità di superficie.

In un ologramma binario la densità di informazione è essenzialmente il numero di "pixel" per unità di area e pertanto il prodotto spazio-ampiezza di banda è semplicemente il numero totale di "pixel" presenti nell'ologramma.

In un ologramma registrato per via ottica il prodotto spazio-ampiezza di banda è tipicamente dell'ordine di 10^{10}.

Gli ologrammi digitali da noi generati hanno raggiunto per ora un prodotto spazio-ampiezza di banda pari a 10^7 che è ben al di sotto dei limiti imposti dalla tecnologia.

Avendo a disposizione una maggior capacità di calcolo, il nostro metodo di fabbricazione potrebbe servire per generare ologrammi con 10^{11} "pixel".

Altri tentativi di ottenere ologrammi codificati in forma digitale hanno generalmente fatto uso di un tracciatore a penna (plotter) guidato da calcolatore, o dispositivi simili, che producono una rappresentazione fortemente ingrandita della figura di interferenza.

Il disegno è poi riportato in scala mediante una riduzione fotografica.

Il processo ha diverse limitazioni.

Anzitutto il prodotto spazio-ampiezza di banda dei tracciatori grafici con la massima risoluzione è solo di pochi milioni di "pixel" e, inoltre, la precisione con cui i pixel possono essere posizionati nell'intera area dell'ologramma è modesta.

4. Fabbricazione con fascio elettronico

La nostra alternativa al tracciamento in scala ingrandita e alla fotoriduzione è la registrazione diretta dell'ologramma in scala 1:1 sul supporto stesso dell'ologramma. Lo strumento che consente di realizzare questa operazione è un sistema litografico a fascio elettronico, originariamente concepito e sviluppato per disegnare le figure che definiscono gli elementi dei circuiti microelettronici nei circuiti integrati.

Un sistema litografico a fascio elettronico è strettamente affine al microscopio elettronico a scansione. Un fascio di elettroni, in una camera a vuoto, è accelerato da una tensione elevata e poi focalizzato su un piano di lavoro. Al punto focale il fascio elettronico ha un diametro di circa 0,1 micron. La direzione del fascio è guidata da un calcolatore di controllo che applica opportuni segnali a un gruppo di bobine di deflessione. La massima deflessione utile è dell'ordine del millimetro, ma si possono realizzare figure este-

se su un'area più ampia muovendo la slitta su cui è montato il piano di lavoro.

Il supporto su cui il fascio elettronico disegna la figura olografica è un altro prodotto tratto dalla industria microelettronica: una lastra piana di vetro ricoperta su una faccia con un film sottile di cromo metallico e sopra di questo uno strato di polimero chiamato "resist". Nella produzione dei circuiti integrati la lastra di vetro è un supporto intermedio dell'immagine - esso serve come "maschera" per trasferire una figura su una fetta di silicio - ma nel nostro caso essa costituisce il prodotto finale.

La lastra è collocata sul percorso del fascio elettronico che viene guidato su tutte le aree della lastra che devono risultare trasparenti nell'ologramma finito.

In queste aree, impressionate dal fascio elettronico, il resist viene alterato chimicamente e può essere rimosso con un solvente, lasciando esposto il cromo sottostante.

Il cromo esposto, a sua volta viene rimosso mediante attacco chimico.

I nostri primi ologrammi generati da calcolatore furono prodotti con il Sistema di Esposizione a Fascio elettronico Honeywell (EBES), il risultato di uno sviluppo interno iniziato nel 1974.

Il sistema aveva un potere risolvente adeguato alla generazione di figure sufficientemente fini per un impiego in olografia. Il potere risolvente, tuttavia, non è il solo criterio per valutare un dispositivo di tracciamento di ologrammi. È essenziale non solo che le frange individuali vengano accuratamente disegnate, ma altresì che la accuratezza di posizionamento delle stesse venga mantenuta per tutta l'area.

Supponiamo che l'errore medio di posizionamento di una frangia sia dell'1%. Se gli errori sono casuali e incorrelati, la conseguenza è solo una riduzione marginale della nitidezza nelle prestazioni ottiche. Ma se l'errore è sistematico, sempre nella stessa direzione, allora su una distanza equivalente a 100 frange di interferenza l'errore cumulativo di posizionamento eguaglia l'intero passo tra due frange contigue. La conseguenza è una seria distorsione del fronte d'onda diffratto.

Nei nostri esperimenti iniziali con il sistema Honeywell EBES noi identificammo diverse sorgenti di distorsione.

Gli errori erano più rilevanti quando una figura risultava da immagini giustapposte, ciascuna ottenuta da un campo di scansione distinto. Spesso le linee che attraversavano il confine tra due immagini erano disallineate.

Si potrebbe essere tentati di attribuire queste imprecisioni di giustapposizione alla slitta meccanica usata per muovere il substrato da un campo di scansione al successivo.

Tuttavia il movimento della slitta è controllato da un interferometro che misura gli spostamenti con una precisione dell'ordine di una piccola frazione della lunghezza d'onda della luce.

Certamente gli errori meccanici contribuivano alla distorsione, ma altri fattori erano probabilmente molto più importanti: essi comprendevano irregolarità e derive nel sistema elettro-ottico e interferenze dovute a campi magnetici esterni.

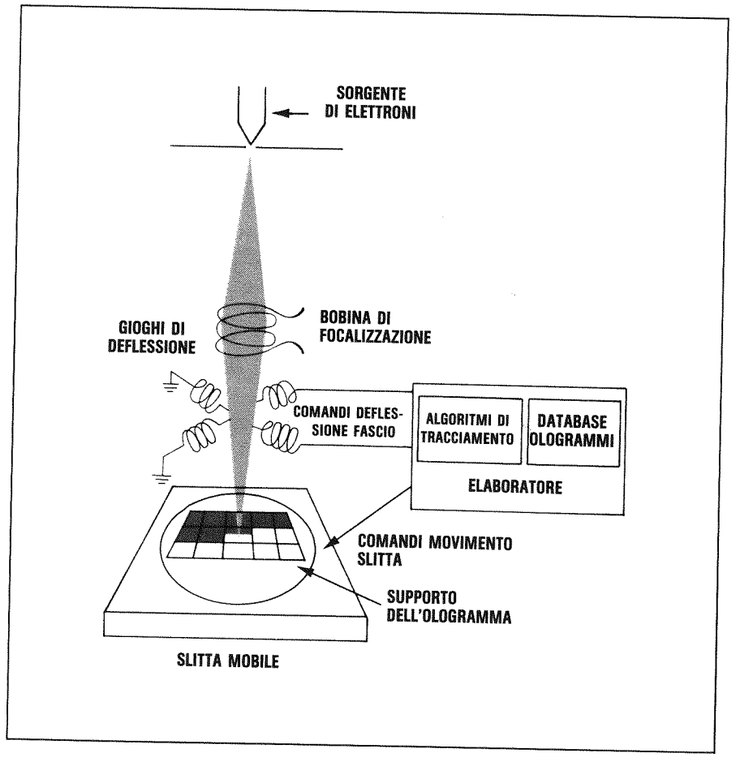


Fig. 2 - Un sistema litografico a fascio elettronico traccia il disegno microscopico dell'ologramma direttamente su un substrato, eliminando passaggi come la fotoriduzione che rendono difficile conseguire la precisione necessaria. Poiché gli elettroni sono particelle dotate di carica elettrica, il fascio di elettroni può essere focalizzato e deflesso da campi magnetici.
La massima deflessione utile del fascio è però dell'ordine di 1 mm.
Ologrammi più grandi sono perciò ottenuti mediante lo spostamento della slitta. Il fascio elettronico è deflesso in corrispondenza di una piccola area del substrato chiamata area di scansione. Dopo che un'area di scansione è stata esposta la slitta viene spostata per portare un altro campo di scansione sotto il fascio elettronico.

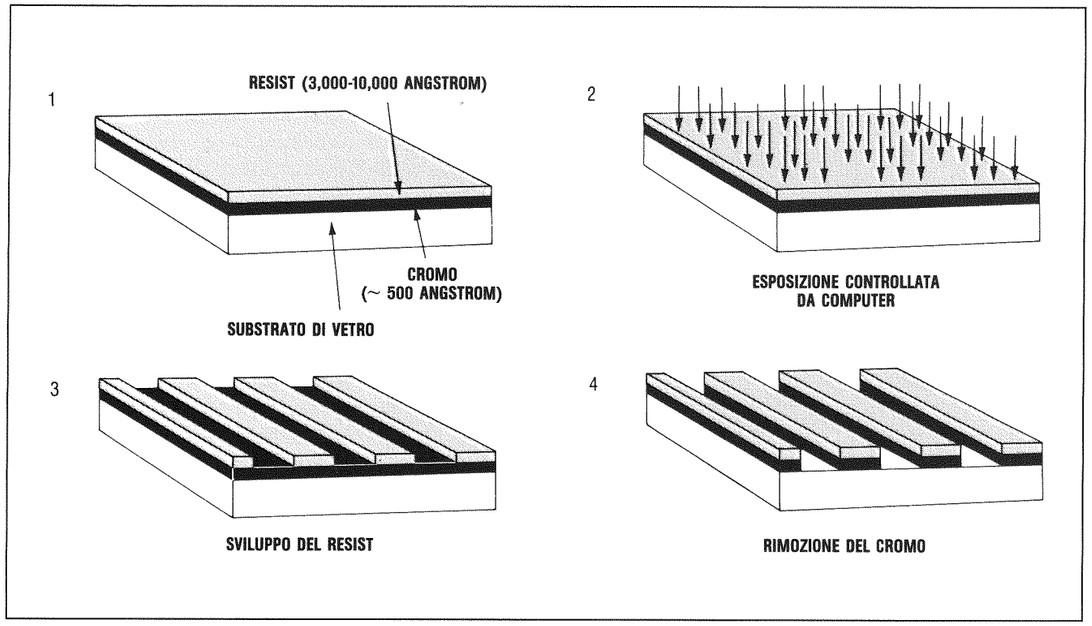


Fig. 3 - Il processo di fabbricazione di un ologramma generato da elaboratore parte da una lastra di vetro coperta da un film sottile di cromo, ricoperto a sua volta da uno strato di "resist". Il sistema di litografia a fascio elettronico espone alcune aree del resist; le aree esposte sono poi rimosse selettivamente con un solvente. La configurazione residua di resist, proteggendo le aree sottostanti, serve come maschera per l'attacco chimico del cromo. Le strisce di cromo rimasto sulla lastra costituiscono le frange dell'ologramma.

Nel 1980 un nuovo dispositivo capace di risoluzione e precisione ancor maggiore si rese disponibile presso la "Solid State Electronics Division". Si tratta di un Microfabbricatore a Fascio Elettronico (EBMF) prodotto dalla Cambridge Instruments.

Lo abbiamo provato con campi di scansione fino a 4x4 mm, ma i migliori risultati sono stati conseguiti per giustapposizione di immagini, ciascuna ottenuta con una campo di scansione limitato a 1 mm.

Anche senza accorgimenti speciali, come compensazione via software, l'errore di posizionamento era contenuto entro ± 0.2 micron ovvero 0.004% su un campo di 1 cm. A partire dal 1981, la maggior parte del nostro lavoro è stata svolta utilizzando questo strumento.

La distorsione totale del sistema era di circa ± 1 micron ossia 0,02% su un ologramma delle dimensioni di 1 cm di lato.

Diversi accorgimenti furono messi in opera per migliorare la precisione del sistema. Il disturbo dei campi magnetici esterni fu eliminato mediante schermatura dell'apparato. Errori sistematici (e perciò riproducibili) del sistema elettro-ottico furono corretti introducendo cambiamenti di compensazione nel programma che controlla la direzione del fascio elettronico.

Gli effetti di fluttuazioni casuali nel sistema magneto ottico furono limitati riducendo le dimensioni del campo di scansione. Con questi sforzi la distorsione fu ridotta a ± 0,15 micron in una direzione di scansione e a ± 0,3 micron nella direzione ortogonale alla precedente ossia approssimativamente al 0.005% su un ologramma delle dimensioni di 1 cm di lato.

5. Codifica della configurazione

La precisione della apparecchiatura di registrazione non è il solo fattore che limita le prestazioni degli ologrammi generati al calcolatore. Al contrario, l'esigenza di elaborazione e memorizzazione di dati costituisce un vincolo ancora più pesante.

Il prodotto spazio-ampiezza di banda ottenibile nei nostri ologrammi è limitato dal tempo necessario per calcolare e codificare una figura di interferenza e dalla quantità dei dati rappresentativi dell'immagine.

L'approccio matematico al calcolo del fronte d'onda è semplice e lineare. Per gli ologrammi che interessano nelle nostre applicazioni è necessario considerare solo come varia la fase della luce in funzione della posizione; l'ampiezza dell'onda elettromagnetica è la stessa dovunque. La natura del calcolo può essere illustrata per mezzo di una figura di interferenza particolarmente semplice, in particolare quella formata dalla intersezione di un fronte d'onda sferico con un fronte d'onda piano, parallelo alla superficie olografica.

Un fronte d'onda sferico è quello generato da una sorgente puntiforme e un fronte d'onda piano è quello generato da un fascio luminoso perfettamente collimato.

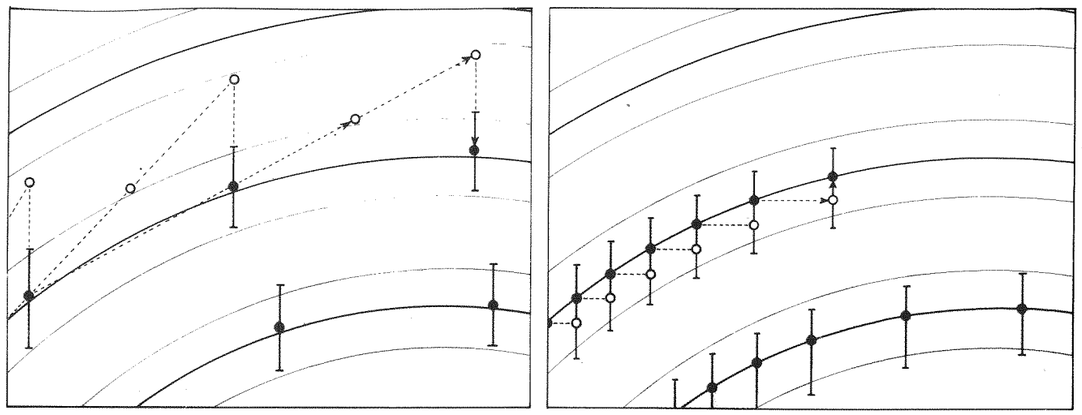


Fig. 4 - L'elaboratore codifica la configurazione di frange in un formato adatto per un dato sistema di litografia a fascio elettronico. La linea centrale di ogni frangia di interferenza (linea curva spessa) è definita come la linea in cui il fronte d'onda richiesto e un fronte d'onda di riferimento sono esattamente in fase. I confini di una frangia (linee curve, sottili) sono definiti come le linee in cui i due fronti d'onda sono sfasati di 90 gradi. Idealmente il fascio elettronico dovrebbe esporre solo le aree contenute entro questi confini. Tuttavia i sistemi a fascio elettronico richiedono che le frange siano costruite partendo da alcune forme primitive. I sistemi di tipo EBES (a sinistra) possono usare trapezoidi con un orientamento qualsiasi. Il sistema Cambridge EBMF (a destra) può usare solo rettangoli allineati con il sistema di coordinate. L'elaboratore usa lo stesso metodo per definire ambedue le primitive. Esso valuta la funzione di fase per identificare un punto che sarà usato per tracciare la linea centrale di frangia (punto pieno). Poi calcola la relazione di fase in corrispondenza di uno o più punti addizionali (circoli) in modo da stabilire l'esatta ampiezza di frangia nonché l'ubicazione del successivo punto della linea centrale.

Supponiamo che l'onda proveniente dalle due sorgenti sia esattamente in fase nel punto in cui il fronte d'onda sferico incontra per primo la superficie olografica. Allontanandosi da questo centro di simmetria, la fase delle due onde deve cominciare a differire perchè la distanza tra la sorgente puntiforme e il piano olografico aumenta. Prima le onde sono esattamente in controfase; poi, quando l'aumento della distanza eguaglia la lunghezza d'onda, esse sono nuovamente in fase. La figura olografica che ne risulta è una macchia centrale, circondata da cerchi concentrici che diventano progressivamente più fini e ravvicinati tra loro.

Da questa analisi si può concludere che il calcolo di una figura di interferenza è interamente un problema geometrico.

Se la forma dei fronti d'onda può essere espressa in forma matematica, allora la fase relativa può essere determinata in ogni punto del piano olografico come funzione delle coordinate del punto.

Il modo più lineare di impostare il calcolo sarebbe di valutare la funzione di fase nella posizione di ogni pixel, poi di assegnare al pixel un valore 1 se le onde sono sfasate meno di 90° e in caso contrario il valore 0.

I valori binari potrebbero poi essere usati per modulare il fascio elettronico, "accendendolo" ad ogni 1 e "spegnendolo" ad ogni 0.

In questo modo la figura di interferenza calcolata sarebbe riprodotta con la massima risoluzione possibile sulla lastra olografica.

Il problema, con questa strategia, è che essa richiede un uso esorbitante di risorse di calcolo. Un ologramma di 1 cm, con pixel ciascuno delle dimensioni di 0,1 micron, contiene più di dieci milioni di pixel, e la funzione di fase dovrebbe essere calcolata più di dieci milioni di volte. Se anche la funzione non è affatto complicata, ciò può richiedere diversi giorni di lavoro di un grosso sistema di elaborazione dati. Inoltre i dati di codifica della figura sarebbero più di dieci miliardi e risulterebbe scomodo e costoso memorizzarli.

La nostra soluzione a questo problema fu di sviluppare uno schema di codifica diverso.

Il primo passo della nostra procedura è la risoluzione dell'equazione di fase allo scopo di identificare i punti in cui le onde sono esattamente in fase; l'insieme dei punti forma delle linee (in generale curve), ciascuna delle quali definisce il centro di una frangia di interferenza. La linea centrale di ogni frangia è poi approssimata da una successione di segmenti rettilinei. La lunghezza di ogni segmento è scelta in funzione della curvatura delle frange: occorre usare segmenti tanto più corti quanto maggiore è la curvatura.

La versione originale del nostro algoritmo fu sviluppata per il sistema Honeywell EBES, che era progettato per esporre in maniera efficiente dei trapezoidi con qualsiasi orientamento. Pertanto il passo finale della procedura era quello di espandere trasversalmente ogni segmento per formare un trapezoide con una ampiezza corrispondente ad uno sfasamento di 90° da ambo i lati della linea centrale. I dati forniti al sistema Honeywell EBES consistevano in una lista di tali trapezoidi.

Tipicamente potevamo definire un ologramma consistente in un miliardo di pixel fornendo la posizione di circa 350.000 trapezoidi di esposizione.

L'introduzione del sistema Cambridge EBMF richiese dei cambiamenti di metodo perchè questo strumento può disegnare trapezoidi disposti secondo pochi orientamenti discreti. Dapprima fu tentato un adattamento diretto del metodo usato col sistema EBES: esso offriva un calcolo efficiente del fronte d'onda e una rappresentazione compatta dei dati, ma lo spezzettamento della figura in campi di scansione richiese un dispendio enorme di tempo di calcolo e generò uno straordinario volume di dati.

Un formato a scalini, costituiti interamente da rettangoli con lati paralleli agli assi x e y, si è dimostrato più pratico ed è stato usato nella maggior parte del nostro lavoro. Abbiamo fatto anche esperimenti con un formato a mosaico che rinuncia all'identificazione delle linee centrali delle frange e assegna invece dei valori binari a blocchi rettangolari di dimensioni prefissate.

Poichè il sistema EBMF non è sempre disponibile per il nostro lavoro, abbiamo recentemente adattato il nostro software a un terzo dispositivo a fascio elettronico, il sistema litografico Varian.

Questa macchina è progettata specificatamente per la produzione di maschere. Essa ha una risoluzione alquanto inferiore a quella offerta dal sistema Cambridge EBMF ma è in grado di effettuare un maggior numero di esposizioni nello stesso tempo. Un altro vantaggio del sistema Varian è che, come il sistema Honeywell EBES, può esporre efficientemente trapezoidi con ogni orientamento. Pertanto siamo nella condizione di poter usare il nostro sistema originale di codifica che è molto più efficiente.

Nel nostro modo corrente di lavorare, la codifica dei fronti d'onda e la suddivisione in campi di scansione sono effettuate utilizzando il sistema Honeywell Multics.

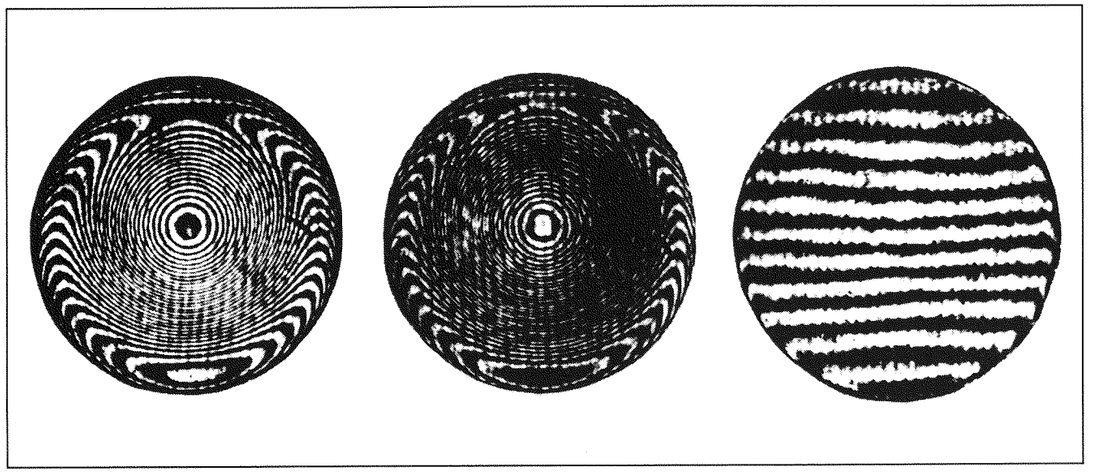
Ciascuno di questi passi richiede tipicamente 10 minuti di elaborazione. A questo punto i dati sono in un formato generico ma sono stati suddivisi per l'impiego con uno specifico sistema a fascio elettronico.

Il passo successivo consiste nella registrazione dei dati su nastro magnetico in un formato accettabile dal sistema grafico a cui sono destinati. Nel caso del sistema Cambridge EBMF non è richiesta alcuna ulteriore elaborazione.

La lastra coperta di resist può essere esposta immediatamente e il processo richiede circa 45 minuti.

Nel caso del sistema Varian i dati devono essere ulteriormente elaborati

Fig. 5 - Gli interferogrammi mostrano come gli ologrammi possono semplificare il controllo di componenti ottici, in special modo quelli che non hanno superfici sferiche. L'interferogramma a sinistra della figura è prodotto da uno specchio parabolico; quello al centro da un ologramma generato mediante elaboratore, che riproduce il fronte d'onda generato da uno specchio parabolico perfetto; e quello a destra è prodotto da un fascio di luce modificato sia dallo specchio che dall'ologramma al centro. Il primo interferogramma è difficile da analizzare, è tuttavia possibile, confrontandolo col secondo interferogramma, identificare le imperfezioni dello specchio. Il terzo interferogramma rappresenta appunto la differenza tra i primi due ed è molto più facile da analizzare. Se lo specchio fosse perfetto, le frange nella terza figura sarebbero rettilinee, parallele ed uniformemente spaziate. L'analisi di questa figura mostra che l'aberrazione media dello specchio (deviazione del fronte d'onda prodotto rispetto a quello desiderato) è pari a 0,03 lunghezze d'onda, ossia inferiore a 0,02 micron.



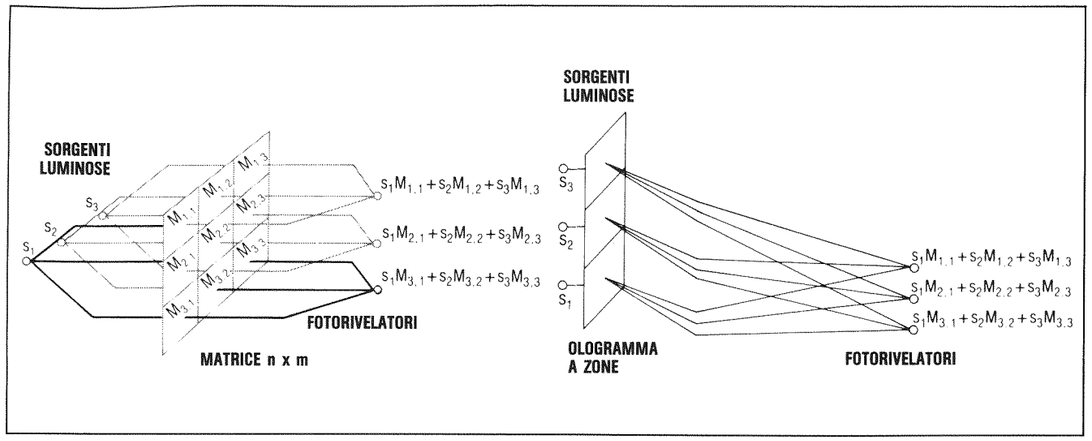


Fig. 6 - Il prodotto di un vettore per una matrice è un calcolo che può essere eseguito otticamente mediante ologrammi. Per eseguire il prodotto vettore matrice i numeri in ogni colonna della matrice devono essere moltiplicati per un elemento corrispondente del vettore e questi valori devono essere poi sommati per righe per definire gli elementi del vettore prodotto.
Un approccio ottico convenzionale (a sinistra) fa uso di una maschera in cui i valori numerici sono rappresentati da aperture di dimensioni diverse (o da differenti toni di grigio).
Complicati sistemi ottici sono però richiesti per convogliare la luce delle diverse sorgenti fino alle aperture appropriate della maschera e per concentrare la luce trasmessa dalle diverse zone sugli appropriati fotorivelatori.
Nel nostro approccio olografico (a destra), ogni sorgente di luce illumina un ologramma che rappresenta una colonna della matrice. L'ologramma è diviso in gruppi di sottozone, ciascuna delle quali diffrange la luce verso uno dei fotorivelatori.
L'area totale di un gruppo di sottozone determina la quantità di luce che viene diffratta e quindi il suo valore numerico. Il passo di frangia definisce l'angolo di diffrazione. Una semplice lente è sufficiente per focalizzare sui fotorivelatori la luce emergente dall'ologramma

per generare una figura di reticolo per ogni campo di scansione.

Questa elaborazione è effettuata da software che gira sul minicalcolatore di comando del sistema Varian.

Il fascio elettronico del Varian espone un campo mediante scansione avanti e indietro, intanto che la superficie di lavoro è mossa trasversalmente sotto il pennello elettronico. La figura di ologramma è generata accendendo e spegnendo il fascio elettronico. Nel sistema Varian una esposizione richiede tipicamente pochi minuti.

6. Verifica ottica

Nella prova di componenti ottici l'utilità di ologrammi generati mediante calcolatore consiste principalmente nella loro capacità di creare fronti d'onda di ogni geometria specificata.

Una lente o uno specchio possono essere considerati come dispositivi per trasformare un fronte d'onda in un altro; una comune lente convessa, per esempio, converte un fronte d'onda piano in un fronte convergente.

Oggi nei sistemi ottici sta diventando sempre più comune includere elementi con forme complesse, in parte perchè il supporto fornito dal calcolatore ha reso meno arduo il loro progetto. La prova dei dispositivi finiti, per verificare che si conformano al progetto non è stata semplificata in pari grado. Lo scopo della prova è di misurare la discrepanza tra il fronte d'onda generato e quello desiderato in base al progetto. Il confronto può essere particolarmente difficile se le superfici delle lenti non hanno una semplice simmetria sferica.

Gli ologrammi generati mediante calcolatore possono essere utilizzati per prove ottiche in diversi modi.

L'approccio più ovvio è di realizzare un ologramma che incorpora il fronte d'onda desiderato, che può poi essere confrontato con il fronte d'onda effettivamente prodotto da un dispositivo ottico per mezzo di una figura di interferenza.

I nostri esperimenti di prova ottica sono stati effettuati con l'interferometro Honeywell/Tropel, uno strumento progettato specificatamente per questo scopo.

La luce, nello strumento, segue un percorso complicato ma l'effetto conclusivo è comunque quello di far passare due fasci di luce attraverso l'ologramma: un fascio viene modificato dal dispositivo in prova e l'altro costituisce un fascio non modificato di riferimento. La porzione del primo fascio che non viene diffratta dall'ologramma è allora combinata con la parte diffratta del fascio di riferimento per produrre una figura di interferenza.

Se il dispositivo in prova è perfetto, il fascio di prova e il fascio di riferimento saranno esattamente in fase e non si verificherà alcuna frangia di interferenza. Una leggera deflessione di un fascio introdurrà allora delle frange rettilinee, parallele e uniformemente spaziate.

Uno dei dispositivi da noi provati era uno specchio parabolico di quattro pollici di diametro. Da una descrizione matematica del paraboloide si è calcolata la figura di interferenza prodotta nel piano dell'ologramma dall'interferenza di un'onda di prova perfettamente conforme al progetto con un fronte d'onda di riferimento. Volendo, sarebbe stato possibile tener conto di imperfezioni note del sistema ottico. La figura di interferenza fu poi codificata e si realizzò un ologramma binario.

Lo strumento Honeywell/Tropel può essere allestito per evidenziare figure di interferenza generate da varie combinazioni di fasci.

Una figura di interferenza basata sul solo specchio parabolico, non influenzata dall'ologramma, è difficile da analizzare perchè la figura ottenuta si scosta in maniera rilevante dalla superficie sferica che darebbe luogo a frange uniformi. Una figura di interferenza basata sull'ologramma ottenuto da calcolatore simula l'effetto prodotto da un paraboloide ideale e può servire come base di confronto. Ancor più utile è la figura di interferenza ottenuta combinando il fascio riflesso dallo specchio parabolico e il fascio di riferimento diffratto dall'ologramma.

Le irregolarità della figura così ottenuta danno una misura diretta delle imperfezioni dello specchio. Nel nostro caso abbiamo verificato che la deviazione dalla superficie ideale era contenuta nel 3% della lunghezza d'onda usata.

Un compito ancor più difficile è verificare un componente ottico che non solo è asferico, ma che in aggiunta non ha simmetria centrale rispetto a un asse.

Abbiamo simulato questo componente combinando uno specchio sferico concavo con uno specchio piano oscillante.

Per creare un ologramma per l'inconsueta figura di interferenza prodotta da questo sistema sono state necessarie solo piccole modifiche al nostro programma. La routine standard impiegata non era in grado di descrivere il fronte d'onda così generato. Tuttavia una volta prodotto l'ologramma la verifica è stata immediata e il successo dell'esperimento ha dimostrato che gli ologrammi generati da calcolatore e realizzati con un fascio elettronico sono in grado di verificare qualsiasi superficie asferica.

7. Elaborazione Ottica

Esiste una attraente simmetria nel concetto di usare ologrammi generati da calcolatore per realizzare dispositivi di calcolo. Dispositivi di questo tipo sono stati presi in considerazione da qualche tempo perchè sono potenzialmente suscettibili di alte velocità di calcolo e particolarmente adatti a metodi di calcolo parallelo (in cui più operazioni sono effettuate simultaneamente). Un'altra area di possibili applicazioni è l'elaborazione di immagini, dove l'informazione è già essenzialmente in forma ottica.

Noi abbiamo investigato l'uso di ologrammi generati con fascio elettronico in un dispositivo ottico specializzato per una operazione matematica particolare: la moltiplicazione di un vettore e una matrice. In questo contesto il vettore è una struttura matriciale monodimensionale, o, in altre parole, una lista di elementi numerici. Il prodotto di un vettore x e di una matrice M si ottiene moltiplicando ogni colonna di M per un elemento di x. I prodotti della moltiplicazione devono poi essere sommati per righe per fornire un elemento del vettore prodotto y.

Il moltiplicatore ottico da noi progettato è utile nel caso in cui vettori variabili in ingresso devono essere tutti moltiplicati per la medesima matrice costante.

È facile concepire uno schema ottico per la moltiplicazione di vettori.

Un vettore moltiplicando di n elementi può essere rappresentato da una successione allineata di n sorgenti di luce, e un vettore prodotto di m elementi può essere rappresentato da una successione allineata di m fotorilevatori.

La matrice costante di moltiplicazione può essere realizzata come una maschera a matrice di nxm celle.

Un sistema ottico può scomporre ognuno dei possibili n fasci di luce in ingresso in m fasci di luce, ciascuno incidente su una delle n celle di una colonna della matrice. Ogni cella trasmette una frazione della luce incidente, determinata dal valore numerico della cella. Un secondo sistema ottico può poi concentrare la luce trasmessa da ogni cella a un fotorivelatore prestabilito.

Sebbene lo schema sia concettualmente semplice, la sua realizzazione pratica non è per niente agevole. Il sistema ottico risulterebbe particolarmente complesso e la scomposizione dei fasci di luce per ottenere una illuminazione uniforme della maschera sarebbe assai difficile. Ne conseguirebbero un campo di lavoro e una precisione assai modesta.

Il nostro approccio sostituisce la maschera a nxm celle con una riga di n ologrammi generati da calcolatore, ognuno degli ologrammi essendo suddiviso in m zone superficiali. Ogni ologramma è illuminato da un singolo fascio di luce ma le sue zone superficiali diffrangono la luce verso ciascuno degli m fotorivelatori.

La quantità di luce diffratta da una zona superficiale prefissata è proporzionale ad un valore numerico associato alla zona.

Un moltiplicatore costruito in questo modo risulta semplice dal punto di vista ottico ed efficiente. Inoltre il processo di litografia a fascio elettronico offre la prospettiva di un elevato prodotto spazio-ampiezza di banda e, conseguentemente, valori di campo di lavoro numerico e precisione elevati.

La figura di interferenza in ogni zona dell'ologramma è una semplice griglia lineare, costituita da linee rettilinee ed equispaziate. La spaziatura fra le frange determina l'angolo di diffrazione. Se l'illuminazione è uniforme su tutto l'ologramma, la quantità di luce diffratta da una zona determinata dipende solo dall'area della zona. Perciò il disegno dell'ologramma richiede solo la definizione dei passi di griglia necessari per dirigere la luce sui fotorivelatori appropriati e l'attribuzione a ogni zona di un'area proporzionale al valore numerico dell'elemento matriciale corrispondente.

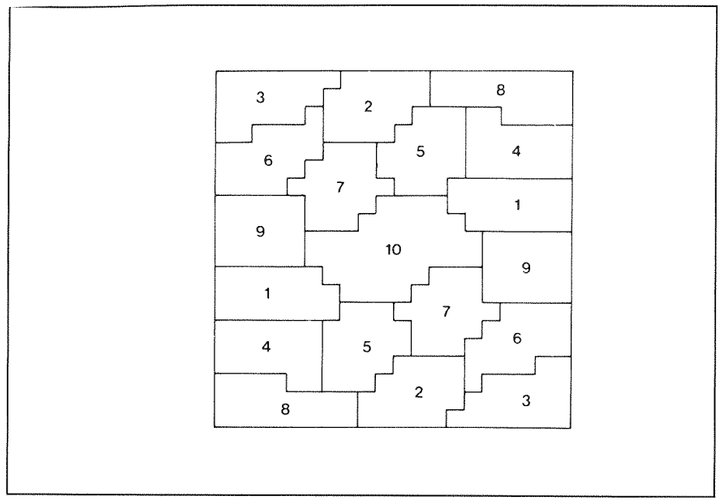


Fig. 7 - Questo moltiplicatore olografico di matrice rappresenta una colonna di una matrice in cui vi sono dieci valori eguali. Ogni gruppo di sottozone ha la stessa area degli altri gruppi: se l'ologramma fosse illuminato uniformemente produrrebbe dieci fasci luminosi di equale intensità. I numeri nelle aree indicano i fotorivelatori verso i quali le diverse aree diffrangono la luce. Sottozone con lo stesso numero hanno lo stesso passo di frangia.
La ripartizione dell'ologramma è stata ottenuta con l'aiuto di un algoritmo diretto a stabilire il miglior compromesso tra diversi requisiti, compresa la simmetria della figura per ridurre gli errori causati da spostamenti irregolari del fascio luminoso incidente.

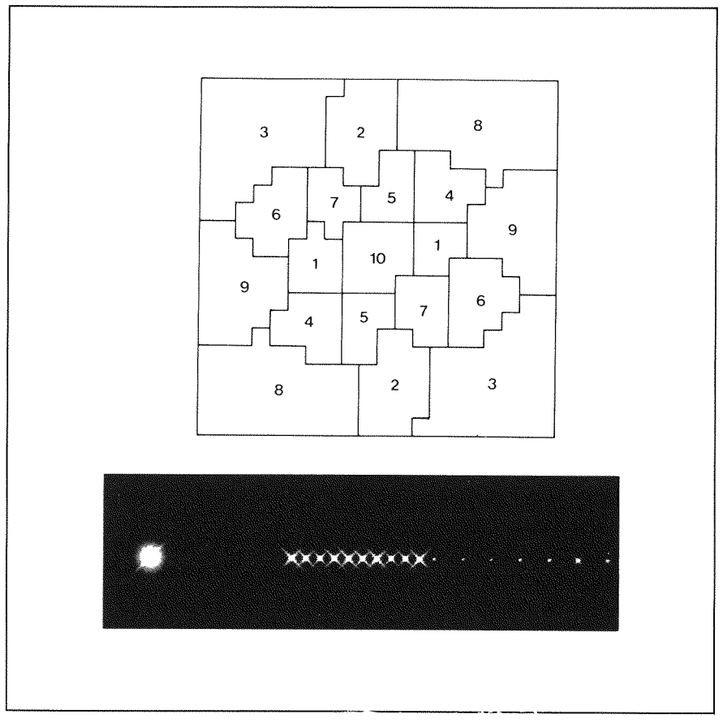


Fig. 8 - Una distribuzione gaussiana di illuminazione impone di tener conto sia della posizione sia dell'area delle diverse sottozone nella ripartizione dell'ologramma. Questa ripartizione dovrebbe produrre dieci fasci di eguale intensità luminosa se gli angoli dell'ologramma fossero illuminati con una intensità pari al 6% dell'intensità di illuminazione delle zone centrali. Nella figura di diffrazione prodotta dall'ologramma (figura inferiore) le dieci immagini chiare al centro corrispondono a fasci diffratti del primo ordine.
I lobi laterali brillanti di ogni fascio appaiono con angolazione di 45° rispetto alla linea dei rivelatori perché le frange sono state tracciate intenzionalmente in diagonale entro ogni sottozona dell'ologramma. Se decorressero verticalmente od orizzontalmente, il fascio diffratto apparirebbe come un + anziché una x, e una piccola frazione della luce di ogni fascio cadrebbe sui fotorivelatori adiacenti.
La macchia brillante a sinistra è una immagine del fascio non diffratto.
Le macchie a destra costituiscono fasci di secondo ordine.

La ripartizione di un ologramma in zone costituisce un interessante problema di calcolo combinatorio. Per sopprimere la diffrazione indesiderata prodotta dalla figura costituita dalle zone dell'ologramma, è necessario stabilire una dimensione minima di zona ed è conveniente realizzare le diverse zone come multipli di questa sottozona elementare. Noi abbiamo scelto una dimensione minima quadrata di 0,5 mm di lato, cosicchè un ologramma quadrato di 1 cm di lato può essere considerato come una griglia di 400 sottozone. Il problema a questo punto è quello di assegnare ciascuna delle 400 sottozone a uno degli m canali di diffrazione.
Per un ologramma che deve scomporre un fascio di ingresso su 10 canali di uscita ci sono 10^{400} possibilità di combinazione!

Come fare per stabilire quale è la combinazione ottima?

Per la ripartizione abbiamo sviluppato quattro criteri. Anzitutto, la luce incidente sulle sottozone assegnate a un canale deve essere proporzionale al valore numerico assegnato alla zona. In secondo luogo le sottozone di una zona devono essere connesse e costituire cioè un'area unitaria.

Gli altri due criteri vengono utilizzati quando il fascio che illumina l'ologramma non è uniforme ma ha una distribuzione gaussiana con maggiore intensità al centro e minore intensità ai bordi. Conviene in questo caso disporre le zone associate ai valori numerici più bassi, vicino ai bordi, in modo da realizzarle con superfici più ampie. È anche desiderabile disporre le subzone di ciascuna zona simmetricamente rispetto al centro dell'ologramma (in contrasto con il secondo criterio) per ridurre al minimo gli errori provocati da possibili escursioni del fascio incidente. In generale i quattro criteri non possono essere rispettati tutti insieme. Per esempio una zona non può essere simmetrica rispetto al centro, collocata vicino ai bordi e al tempo

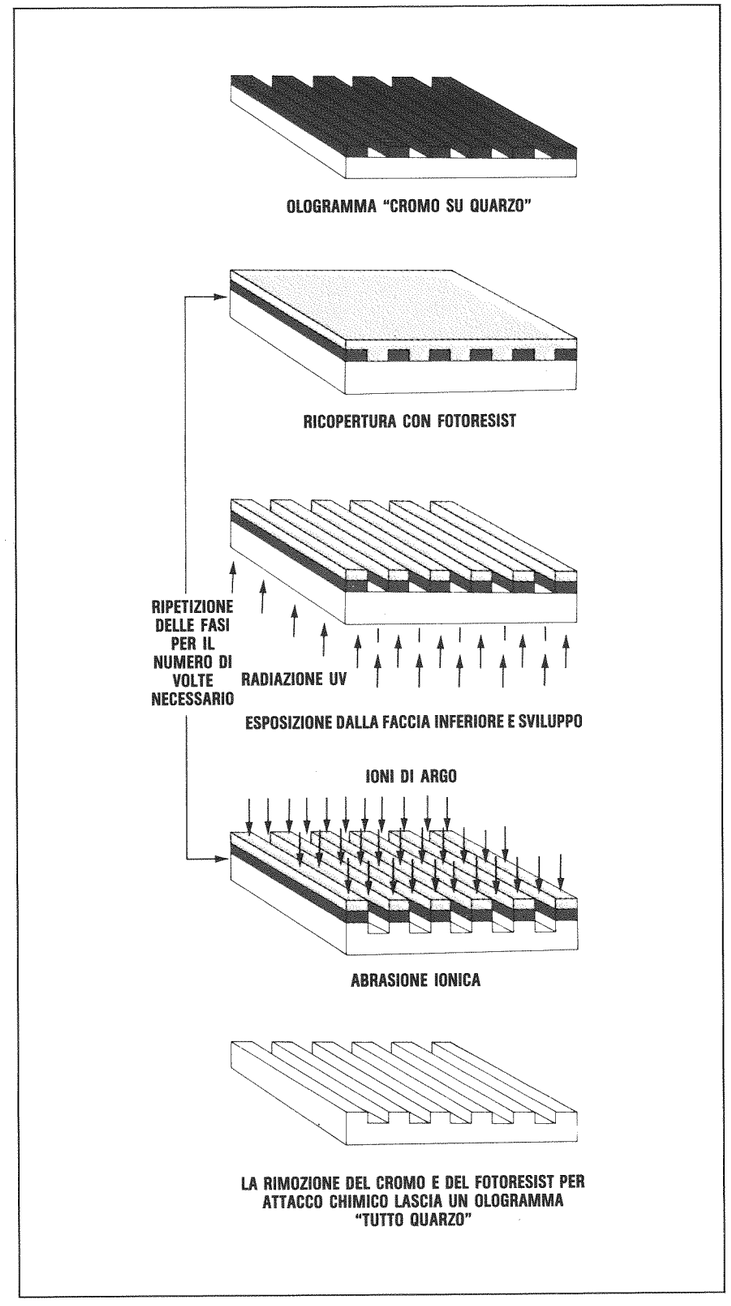


Fig. 9 - L'abrasione ionica può essere usata per produrre ologrammi che generano fasci di uscita più luminosi.
Un ologramma cromo-su-quarzo è coperto con un polimero, ("fotoresist"), sensibile ai raggi ultravioletti.
Il cromo agisce come maschera, proteggendo certe zone del fotoresist quando la faccia inferiore della lastra è esposta a luce ultravioletta.
Il fotoresist esposto alla luce viene poi rimosso e la faccia superiore della lastra è poi bombardata con ioni di argo.
Gli ioni rimuovono il quarzo ad una velocità doppia rispetto alla velocità di rimozione del fotoresist ed è possibile incidere dei solchi di profondità considerevole con ripetute mascherature di fotoresist. Nell'ultimo passo del processo costruttivo il fotoresist ed il cromo vengono rimossi dalla lastra.
L'ologramma finito è un elemento tutto-vetro, in cui le zone più spesse svolgono la funzione delle frange.
Mentre gli ologrammi in cromo su vetro lavorano bloccando la luce che interferirebbe negativamente con i fasci diffratti, le creste di vetro dell'ologramma molato-a-ioni inducono una variazione di fase di 180° della luce, in modo che questa interferisce additivamente con la luce trasmessa attraverso i solchi.

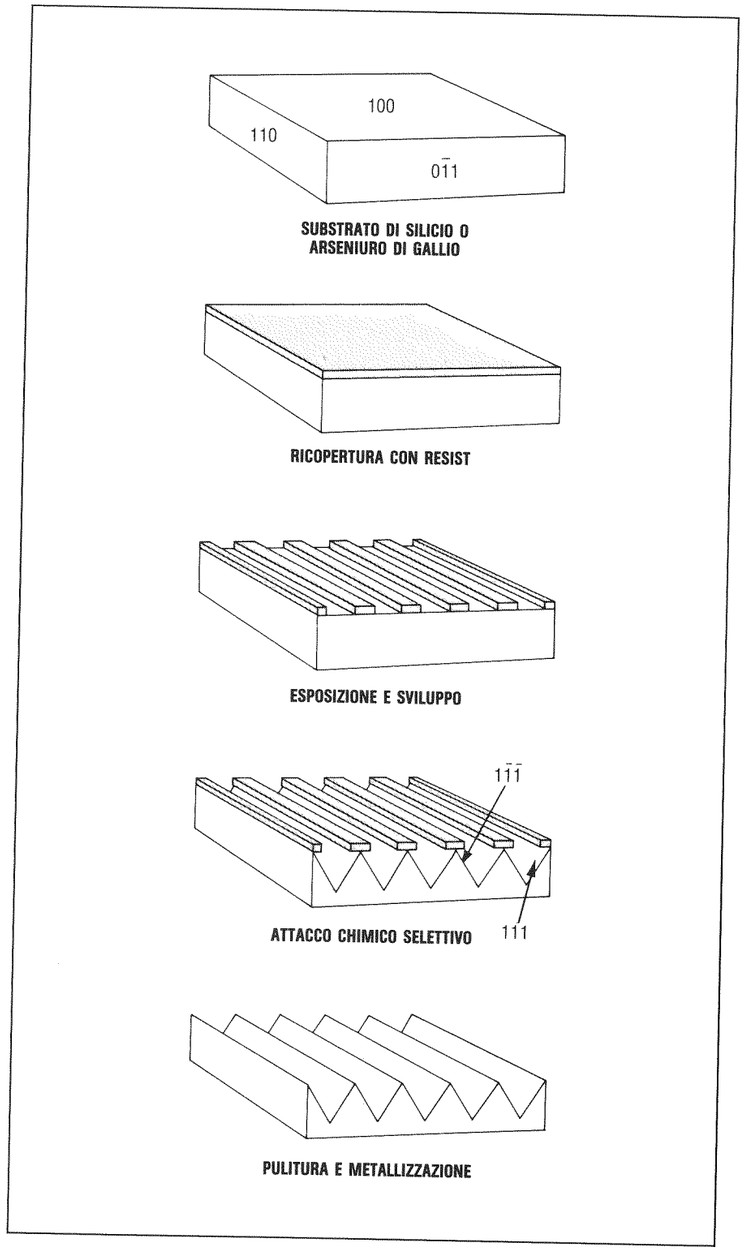


Fig. 10 - Gli ologrammi incisi potrebbero produrre fasci di uscita ancora più luminosi degli ologrammi ottenuti per abrasione ionica. Qui il substrato è costituito da una lastra di silicio monocristallino anziché da vetro.
La chiave del processo è un agente di attacco anisotropico che rimuove più velocemente il materiale in direzione di alcuni assi del reticolo cristallino e più lentamente in altre direzioni.
Un rivestimento di resist è applicato al substrato, previa passivazione con nitruro di silicio.
Strisce di resist, approssimativamente parallele ai piani che nel silicio vengono attaccati più lentamente, sono impressionate dal fascio elettronico di un sistema litografico.
Il resist serve come maschera per l'attacco chimico con nitruro di silicio. A sua volta il nitruro di silicio serve come maschera per l'attacco chimico del silicio sottostante.
L'agente anisotropico agisce rimuovendo le aree esposte e formando dei solchi i cui fianchi sono costituiti dai piani secondo cui l'attacco chimico procede più lentamente.
Il silicio è poi pulito e metallizzato.
Da un punto di vista teorico, gli ologrammi incisi possono essere più efficienti degli ologrammi a inversione di fase perché in essi la variazione di fase è continua anziché limitata a due soli valori.

stesso connessa. Un compromesso è necessario.

Abbiamo quindi sviluppato due algoritmi in cui a questi criteri vengono assegnati dei pesi numerici diversi.

In un caso, l'algoritmo considera tutte le possibili varianti di assegnazione di ogni sottozona a un canale.

Se la variante migliora la ripartizione essa viene adottata.

Il secondo algoritmo consente lo scambio nella assegnazione di due sottozone a un canale. Anche con questo supporto di calcolo però, il ritocco dei valori numerici per la ripartizione delle zone deve essere guidato dal giudizio umano.

Un totale di 400 sottozone consente di rappresentare solo un campo numerico limitato e con una precisione modesta.

Se l'illuminazione è uniforme, il rapporto tra il valore massimo e il valore minimo rappresentabili non può essere superiore a 400 e la risoluzione non può essere inferiore a 1/400 del valore massimo. Tuttavia se l'illuminazione ha una distribuzione gaussiana, il numero utile equivalente di sottozone è maggiore di 5000. In questo caso le sottozone agli angoli, che ricevono solamente il 6% circa della luce ricevuta dalle

sottozone disposte al centro, sono le unità di base di conteggio e un valore elevato e preciso può essere ottenuto combinando relativamente poche sottozone centrali e un conveniente numero di sottozone periferiche.

Il campo e la precisione possono anche essere migliorati mediante "taratura" delle sottozone per regolare la loro uscita.

Il metodo migliore di taratura consiste nell'aggiungere una regione a griglia negativa, ossia una regione in cui le frange di interferenza sono esattamente in opposizione di fase rispetto a quelle esistenti nella sottozona. Ciò è più facile da realizzare di quanto possa sembrare. La regione negativa può essere registrata in un'area già occupata da una sottozona positiva e l'effetto di questa sovrapposizione di frange positive e negative è quello di lasciare interamente scoperta una certa area.

Con disappunto si deve riferire che i nostri esperimenti di moltiplicazione di matrici e vettori per via olografica non hanno conseguito gli obiettivi prefissati. Le nostre prime griglie erano progettate per generare 10 uscite di eguale intensità, quando illuminate uniformemente. Le uscite effettive differivano anche per un fattore prossimo a 2.

La maggior sorgente di errore era dovuta all'illuminazione di fatto non uniforme.

Perciò ci orientammo allo sviluppo di griglie progettate per una illuminazione distribuita secondo un profilo gaussiano.

I risultati furono molto migliori e quando l'apparato era correttamente tarato le uscite differivano non più dell'uno per cento. Tuttavia un altro problema prese consistenza. Per via delle fluttuazioni spaziali del fascio laser, la stabilità dell'apparato giorno dopo giorno non era migliore dell'1%.

Appare dunque chiaro che lo sfruttamento di ologrammi per il calcolo ottico richiederà lo sviluppo di sorgenti di luce molto più stabili di quelle attualmente disponibili.

8. Ologrammi ed alta efficienza

Il maggior inconveniente dei nostri ologrammi di cromo depositato su vetro consiste nella loro bassa efficienza diffrattiva. Solo il dieci per cento circa della luce incidente viene diffratto.

Più della metà della luce incidente è persa per assorbimento dovuto alle frange opache del cromo che copre metà dell'area e la maggior parte della luce rimanente passa attraverso la griglia senza subire deflessione.

Le dispersioni costituiscono un serio problema in molte applicazioni, incluso il calcolo ottico.

C'è tuttavia un potenziale di maggiore efficienza in ologrammi che lavorano non per assorbimento della luce incidente ma per modifica della sua fase. Per esempio se le frange opache nel nostro ologramma fossero sostituite da qualche struttura che trasmette la luce ma ne modifica la fase di 180 gradi, la maggior parte della luce incidente sarebbe disponibile per la generazione del raggio diffratto. Mentre una griglia di assorbimento sottrae componenti selezionate alla luce trasmessa, una griglia di fase somma di fatto l'immagine negativa di queste.

Come si può ottenere un cambiamento di fase?

È noto che ogni qualvolta la luce passa attraverso una lastra di vetro la sua fase viene modificata perchè la velocità della luce è minore nel vetro che non nell'aria (il rapporto delle velocità costituisce l'indice di rifrazione del materiale). Se la lastra di vetro ha uno spessore uniforme, la variazione di fase è la stessa in ogni punto e non vi sono conseguenze significative.

Tuttavia se certe aree sono più spesse di altre, la luce che attraversa le regioni più spesse sarà ritardata di più. Mediante una scelta della differenza di spessore in rapporto alla frequenza della luce, la variazione di fase può essere resa eguale a metà della lunghezza d'onda ovvero a 180 gradi.

La strategia di base per ottenere un ologramma di fase con il nostro ap-

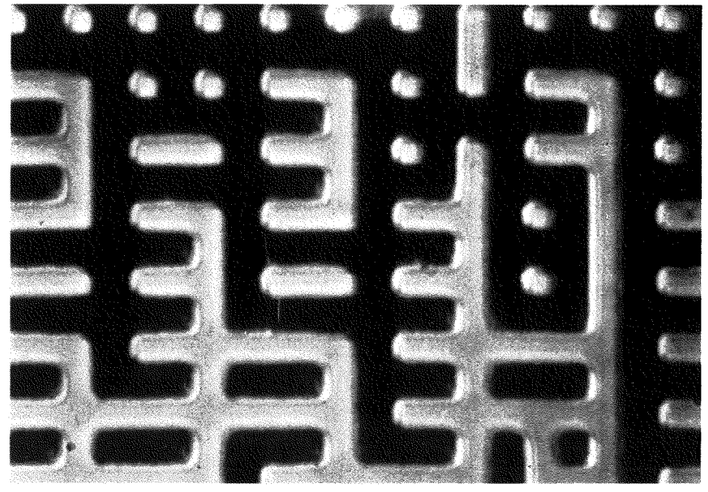


Fig. 11 - Ologrammi per filtraggio spaziale sono impiegati in sistemi ottici per rivelare la presenza di specifiche figure in una immagine.
La fotografia mostra una porzione di ologramma usato per riconoscere dei quadrati.
Se in una immagine è presente un quadrato, l'ologramma trasmette una maggiore quantità di luce. Questo ologramma è stato codificato con un metodo a mosaico che implica la divisione dell'ologramma in celle di uguali dimensioni e il calcolo della funzione di fase al centro di ogni cella. L'ologramma è costituito da 512x512 celle quadrate ciascuna con un lato di 2 micron.

parato a fascio elettronico è estremamente semplice. Anzitutto la figura di ologramma è formata nel modo usuale, come deposito di cromo su vetro. La lastra è poi sottoposta ad un processo di "molatura a ioni," in cui ioni di argo accelerati contro la lastra attaccano il vetro in quelle aree che non sono protette da un rivestimento superficiale. Ne consegue la formazione di solchi a fondo piano tra le frange. Quando i solchi hanno raggiunto la profondità richiesta il cromo viene rimosso lasciando una lastra olografica in cui le frange di interferenza sono definite da rilievi superficiali. Si ottiene così un ologramma durevole, interamente in vetro.

In pratica si sono adottate alcune varianti al procedimento di fabbricazione, ma il principio di base rimane lo stesso.

Per verificare il principio, abbiamo prodotto per abrasione ionica una serie di lenti olografiche progettate per concentrare un fascio di luce collimato in un punto focale. Alcune di queste lenti, con frange di ampiezza pari a 0,5 micron, un passo minimo di frangia pari a 1 micron e 3700 frange, misero alla prova sia le nostre risorse di elaborazione dati sia l'apparato a fascio elettronico.

La massima efficienza teorica di un ologramma a fase binaria è circa del 40% e la maggior parte delle lenti da noi prodotte si approssimò a questo valore.

L'efficienza dovrebbe aumentare quanto più il passo di griglia è piccolo ma difficoltà nel controllo dell'ampiezza dei solchi hanno limitato l'efficienza nel caso migliore a circa il 20%.

Per un ologramma che lavora sulla base di una variazione di fase continua, anzichè binaria, l'efficienza può raggiungere teoricamente il 100%.

Le frange di interferenza in tale ologramma sono formate da solchi a V, tagliati o "tracciati" in una superficie riflettente.

Le griglie di diffrazione ottenute con un dispositivo di tracciatura meccanica sono di questo tipo. Assegnati un profilo ideale di solco, una spaziatura di frangia e l'angolo di incidenza della luce, una griglia tracciata può convogliare praticamente tutta la luce incidente in un fascio diffratto.

Noi abbiamo adattato, allo scopo, il processo di litografia con fascio elettronico per la tracciatura di ologrammi a riflessione su silicio monocristallino.

Il processo comincia con l'impiego di un substrato monocristallino in cui la superficie ha un orientamento noto rispetto agli assi di simmetria cristallografica. Uno strato di nitruro di silicio è deposto sulla superficie e coperto con un film di "resist" per fascio elettronico.

La figura di interferenza è poi disegnata dal dispositivo a fascio elettronico e il nitruro di silicio è rimosso per attacco chimico in corrispondenza delle regioni esposte, lasciando delle sottili strisce di nitruro in corrispondenza della linea centrale delle frange.

Il nitruro residuo serve come maschera per l'ultima fase del processo: attacco chimico anisotropico del silicio. L'agente chimico rimuove il materiale più velocemente in direzione di certi assi del reticolo cristallino e più lentamente in altre direzioni. Come risultato, quando l'agente chimico ha terminato la sua azione, i piani esposti tendono ad essere quelli che vengono corrosi più lentamente. Nei nostri esperimenti abbiamo impiegato un agente chimico che agisce più lentamente sui piani che in notazione cristallografica sono identificati con 111.

Essi formano dei ripidi solchi con pareti a 54,7 gradi rispetto alla verticale.

Idealmente i solchi avranno dei picchi affilati e dei fondovalle formati dall'intersezione di questi piani.

Le griglie tracciate che siamo stati in grado di produrre sono risultate lungi dall'ideale.

Ciò era dovuto all'attacco chimico del silicio, che si estendeva anche sotto la maschera di nitruro, col risultato di produrre delle terrazze anzichè picchi affilati.

Le aree piane della superficie riflettevano il fascio anzichè diffrangerlo, riducendo in misura notevole l'efficienza.

A dispetto di certe contrarietà, tuttavia, si è dimostrata chiaramente l'utilità della litografia a fascio elettronico per la produzione di ologrammi.

I nostri ologrammi sono usati correntemente per la prova di componenti ottici. Il prodotto spazio-ampiezza di banda è per questi di un ordine di grandezza superiore a quello di ologrammi ottenuti con altri metodi e in definitiva esiste il potenziale per conseguire le prestazioni ottenibili da ologrammi realizzati mediante registrazione ottica.

Ma forse, e ancor più importante, la possibilità di realizzare fronti d'onda di conformazione arbitraria rende la luce stessa un mezzo plastico che può essere plasmato secondo l'immaginazione e la fantasia del progettista.

Memorie a prova d'errore

CINZIA ANEDDA,
LEONARDO FELICIAN,
TIZIANA TIZIANEL

Istituto di matematica, Informatica e Sistemistica
Università degli Studi di Udine

1 - Introduzione

Le grandi memorie degli odierni elaboratori, sia quelle centrali che quelle di massa, sono costruite con tecnologie molto affidabili. Si prevede che ogni chip di cui è composta la memoria centrale possa funzionare senza manifestare errori per decine di anni. Il problema è che collegando nella memoria di un grande elaboratore centinaia di chip, si abbassa pericolosamente, fino ad arrivare a poche ore, il tempo necessario perché uno di essi presenti un errore. Un discorso analogo vale per le memoria di massa.

Se si considera l'enorme importanza dei dati contenuti nelle memorie di un calcolatore impiegato per scopi gestionali o scientifici, si comprende perché sia necessario risolvere il problema degli errori mediante strumenti matematici come i codici rivelatori e correttori di errori. Tali codici consentono ai dispositivi di memoria di funzionare perfettamente anche dopo centinaia di errori, trasformando memorie altamente inaffidabili in memoria altamente affidabili. Sembra che questo sia l'unico modo per realizzare grandi memorie con un elevato grado di affidabilità.

In questo articolo si intende spiegare da cosa sono causati gli errori nelle memorie, in correlazione con le loro caratteristiche hardware, presentare il principio di funzionamento dei codici rivelatori e correttori di errore di Hamming e infine valutare brevemente l'affidabilità delle memorie protette da tali codici.

2. Caratteristiche hardware delle memorie e formazione degli errori.

Si esaminano dapprima le cause del verificarsi degli errori nelle memorie dei calcolatori in relazione alle loro caratteristiche hardware.

Va precisato innanzitutto che solitamente la memoria di un calcolatore si divide in:

- memoria centrale (caratterizzata da dimensioni limitate, bassi tempi di accesso, alto costo e costruita con chip di silicio)
- memoria di massa (caratterizzata da notevoli dimensioni, più alti tempi di accesso, costo più basso; le tecnologie di costruzione sono quella magnetica, per dischi e nastri, e quella ottica, solo per dischi)

In entrambi i tipi di memorie, durante il processo di lettura o scrittura (memorizzazione), si può verificare un errore che è sostanzialmente di due tipi.

- *Soft error* (errore logico): si ripete casualmente in posizioni diverse all'interno della memoria. È dovuto alla presenza del "rumore" che si manifesta durante il processo di lettura e dipende dalla capacità di leggere correttamente una data locazione della memoria centrale o di posizionarsi su un preciso punto della superficie di memorizzazione nel caso di dischi e nastri.
- *Hard error* (errore fisico): è dovuto alla rottura o a difetti fisici del supporto di memorizzazione. Gli errori hard si ripetono sempre nella stessa posizione, possono portare ad una totale mancanza di uno o più bit e, al contrario degli errori logici, non vengono riconosciuti con letture ripetute.

2.1. Le memorie centrali

La memoria centrale è costruita a partire da chip di silicio. Un chip di memoria è una matrice quadrata di celle di immagazzinamento di dati. Ad esempio, nel chip da 64 Kbit le celle di memoria sono organizzate sotto forma di matrice quadrata di 256 celle per lato. Ogni cella contiene un bit, cioè una cifra binaria 0 o 1. Il chip è ad accesso casuale cioè il contenuto di ciascuna cella può essere memorizzato ("scritto") o recuperato ("letto") individualmente, poiché ogni cella ha un indirizzo univoco.

Gli 0 e gli 1 sono rappresentati dalla presenza o assenza di una carica elettrica in posizioni, all'interno del cristallo di silicio, le cui proprietà elettriche ne fanno buche di potenziale. Per memorizzare 0 in una cella la buca di potenziale corrispondente viene riempita di elettroni; al contrario per memorizzare 1 essa viene svuotata di elettroni.

Quando si leggono i contenuti, si misura la carica negativa di quella posizione: se la carica supera un certo valore (soglia), il bit viene riconosciuto come 0, altrimenti come 1.

Si possono però verificare errori hard o soft. Gli errori soft sono solitamente dovuti alle particelle alfa (nuclei di elio) emesse dai nuclei atomici pesanti, comuni in queste memorie, durante il decadimento radioattivo.
Quando una particella alfa, dotata di energia elevata, penetra in una buca di potenziale, strappa gli elettroni dalle loro posizioni. Questi possono raggiungere una buca di potenziale vicina e causare un errore se questa è vuota: l'1 prima contenuto viene trasformato in uno 0.

2.2. Le memorie di massa magnetiche

I materiali ferromagnetici contengono un gran numero di piccole regioni (domini di Weiss) in ciascuna delle quali il campo magnetico è orientato in modo casuale. L'applicazione di un campo magnetico esterno provoca l'allineamento totale dei campi magnetici microscopici dei domini nella direzione del campo esterno. Il sistema utilizzato per la memorizzazione delle informazioni consiste dunque nel magnetizzare delle piccole aree su superfici ricoperte da un sottile strato di materiale magnetizzabile, aree che costituiscono le celle di memoria.

L'accesso alle aree avviende mediante scorrimento meccanico uniforme del supporto rispetto ad una testina di lettura-scrittura costituita da un nucleo di materiale ferromagnetico su cui vi è un avvolgimento. Il passaggio di corrente nell'avvolgimento della testina di scrittura crea, nelle vicinanze del traferro, un campo magnetico che magnetizza il supporto sottostante in direzione coerente con il verso della corrente. Sullo strato magnetico dopo la registrazione vi sarà pertanto una sequenza di celle magnetizzate.
Quando il supporto passerà sotto la testina di lettura, vi sarà una variazione di flusso nel nucleo ferromagnetico in corrispondenza delle variazioni di magnetizzazione del supporto. Questa variazione indurrà una forza elettromotrice nell'avvolgimento della testina di lettura; dall'andamento della tensione ai capi dell'avvolgimento si può ricostruire l'andamento della cor rente di scrittura e quindi l'informazione scritta. Le informazioni devono essere, come sempre, codificate come sequenze di 0 e 1; si tratta di inviare quindi sull'avvolgimento della testina di scrittura una corrente che con i suoi cambiamenti possa rappresentare la sequenza di 0 e 1.

Ci sono quattro metodi fondamentali di registrazione (cfr. figura 1):

- *Ritorno a polarizzazione* (RB, Return to Bias): il supporto è magnetizzato permanentemente in una direzione, corrispondente ad esempio al bit 0. Per scrivere un 1 la magnetizzazione viene rovesciata per una parte dell'intervallo assegnato ad un bit e poi riportata a 0 prima dell'inizio del successivo intervallo.

- *Non ritorno a zero* (NRZ, Non Return to Zero): la magnetizzazione è costante durante ciascun intervallo e la sua direzione definisce il valore 1 o 0 del bit.

- *Modulazione di fase* (PM, Phase Modulation): si effettua sempre un'inversione della magnetizzazione a metà dell'intervallo corrispondente a un bit, in modo che l'andamento relativo all'1 risulti in opposizione di fase con quello dello 0.

- *Modulazione di frequenza* (FM, Frequency Modulation): si effettua sempre un'inversione di magnetizzazione all'inizio dell'intervallo elementare. Se il bit vale 1 se ne effettua un'altra al centro dell'intervallo.

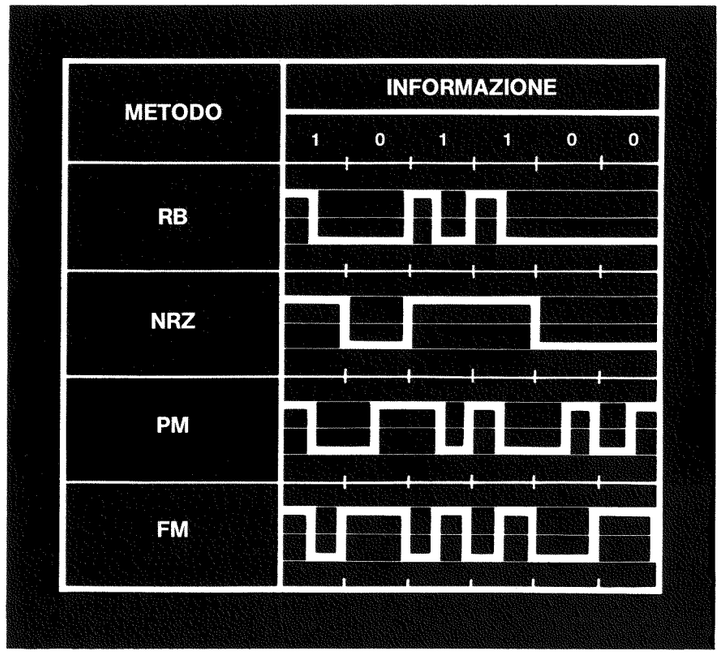


Fig. 1 - Le correnti di scrittura nei 4 metodi di registrazione.

I metodi di trasformazione del segnale analogico (variazione del flusso concatenato con la testina) in un equivalente digitale dipendono dal metodo di memorizzazione usato e influenzano l'affidabilità del processo di lettura e la densità dei dati memorizzati.

Infatti, si può notare che con il metodo NRZ si ha al più un'inversione di magnetizzazione per ciascun bit; quindi, per la stessa densità di inversioni della magnetizzazione, c'è un numero di bit memorizzati doppio rispetto agli altri sistemi illustrati. Per questo il metodo NRZ è il più usato.

Riassumendo, la densità dei dati, che si cerca di tenere alta, è determinata dalla occupazione di spazio di un bit e dalla vicinanza di due bit consecutivi; se un bit è troppo "stretto" o se due bit sono troppo vicini non si riesce più a rivelare le inversioni di magnetizzazione all'interno di un bit o tra due adiacenti. A sua volta, la minima distanza tra due bit è determinata dalla distorsione del segnale di lettura quando i due bit sono troppo vicini. Infatti aumentare la densità significa diminuire la distanza tra i poli magnetici opposti. Una conseguenza di ciò è la tendenza dei poli a neutralizzarsi l'uno l'altro ("self-demagnetization"). In tal caso il campo magnetico diminuisce e la testina di lettura non è più in grado di rilevare le variazioni dello stato di magnetizzazione.

2.3. Le memorie di massa ottiche

Se si adotta la tecnologia di memorizzazione ottica, viene impiegato il laser sia in fase di scrittura sia in fase di lettura. I principali tipi di dispositivi di memorizzazione realizzabili attraverso la tecnologia ottica si possono ripartire in tre classi: *write once,* che possono venir scritti una sola volta dall'utente divenendo poi a sola lettura; *read only,* che possono venir letti ma non scritti dall'utente; *rewritable,* in cui i dati possono venir cancellati e riscritti dall'utente.

Tra le diverse tecniche impiegate per la realizzazione delle memorie write once, quella ablativa è la più utilizzata: in fase di scrittura il laser provoca un foro sulla superficie del disco in corrispondenza di un 1 mentre lascia inalterato il supporto in corrispondenza di uno 0 binario; in fase di lettura viene rilevata la diversa riflettività tra zone forate e zone non forate per determinare la presenza di un 1 o di uno 0.

Per la realizzazione di memorie rewritable sono state sviluppate tecniche che permettono di cancellare e riscrivere i dati, poiché non viene alterato il supporto in modo irreversibile. Queste tecniche sfruttano le caratteristiche dei materiali di cui è costituita la superficie del disco, ad esempio la capacità di passare da uno stato cristallino a uno amorfo oppure di magnetizzarsi, in modo reversibile.

Gli errori nelle memorie ottiche si manifestano soprattutto durante il processo di lettura.

- La tecnica ablativa impone l'esistenza di un forte contrasto tra zone forate (altamente riflettenti) e zone non forate nel materiale che costituisce il supporto di memorizzazione. Infatti un'errata riflettività comporta distorsioni nel segnale di lettura.
- Le tecniche utilizzate per la realizzazione di memorie ottiche cancellabili non sono completamente affidabili sia per eventuali difetti del materiale sia perché dopo un certo numero di cambiamenti di stato si verifica un fenomeno di "fatica" che riduce le prestazioni del supporto. In questo secondo caso la differenza di riflettività tra uno stato e l'altro non è più così accentuata e ciò può causare errori in fase di lettura.

Il processo di scrittura risulta maggiormente protetto: si utilizzano infatti due tecniche di rivelazione degli errori dette *read during write* (RDW) e *read after write* (RAW) che controllano la scrittura dei dati mentre si formano sulla traccia. Per fare ciò, un laser di lettura segue immediatamente quello di scrittura ed è in grado di individuare gli errori non appena si verificano, mediante confronto con il buffer di memoria.

In caso di errore, se è possibile riscrivere (dispositivo rewritabile) si ripete l'operazione, altrimenti (dispositivo write once) il settore contenente dati imperfetti viene contrassegnato come difettoso e i dati vengono riscritti nel settore successivo.

3. Codici rivelatori e correttori

Si sono finora esaminate alcune cause di formazione degli errori nelle memorie, in relazione alle loro caratteristiche hardware. Poiché le informazioni contenute nelle memorie sono un patrimonio di vitale importanza per gli utenti, si comprende perché è necessario individuare e correggere gli errori di memorizzazione che si verificano.

Per fare ciò è indispensabile l'utilizzo di codici rivelatori che sono in grado di segnalare che la parola letta è diversa da quella scritta in memoria; e soprattutto di codici correttori capaci di individuare e correggere i bit che sono stati alterati. L'idea di fondo di questi codici è quella di "allungare" la parola di memoria inserendo dei bit opportuni: in questo modo si rende *ridondante* la parola e si permette l'identificazione e la correzione degli errori.

Il numero di bit aggiuntivi di protezione e l'algoritmo per determinarli in dipendenza dalla parola da memorizzare caratterizzano i vari codici.

Come esempio di codice rivelatore si ricorda il CODICE A CONTROLLO DI PARITÀ: alla fine della parola di memoria viene aggiunto un bit in modo che il numero totale di bit uguali a 1 nella parola prodotta sia pari.

Se la parola da memorizzare è 1101, contenente tre bit a 1, si scrive in memoria 11011 (che contiene un numero pari di bit a 1); analogamente la parola 0101 viene memorizzata come 01010, e così via. In tal modo se in fase di lettura si rileva un numero dispari di bit uguali a 1 si può sicuramente affermare che si è verificato almeno un errore, senza però riuscire ad individuare quale bit è stato invertito.
Si fa notare che se due bit sono stati invertiti, il bit di parità è comunque corretto, e questa situazione di errore non viene rivelata.
Il codice di parità è dunque un codice rivelatore piuttosto debole, tutta-

via non comporta troppa occupazione aggiuntiva di memoria. Per codificare K bit ne bastano K+1.
Si illustrano ora i codici di Hamming, che sono stati i primi codici correttori di un certo rilievo, proposti da Richard Hamming dei Bell Telephone Laboratories nel 1948.
Si descriverà il principio di funzionamento del codice di Hamming in grado di correggere 1 errore e una variante di questo codice che consente anche il rilevamento di 2 errori.

3.1. Il codice correttore di Hamming

Per descrivere i codici correttori di Hamming è necessario precisare quanti bit si devono aggiungere alla parola da memorizzare e descrivere l'algoritmo che ne determina il valore.

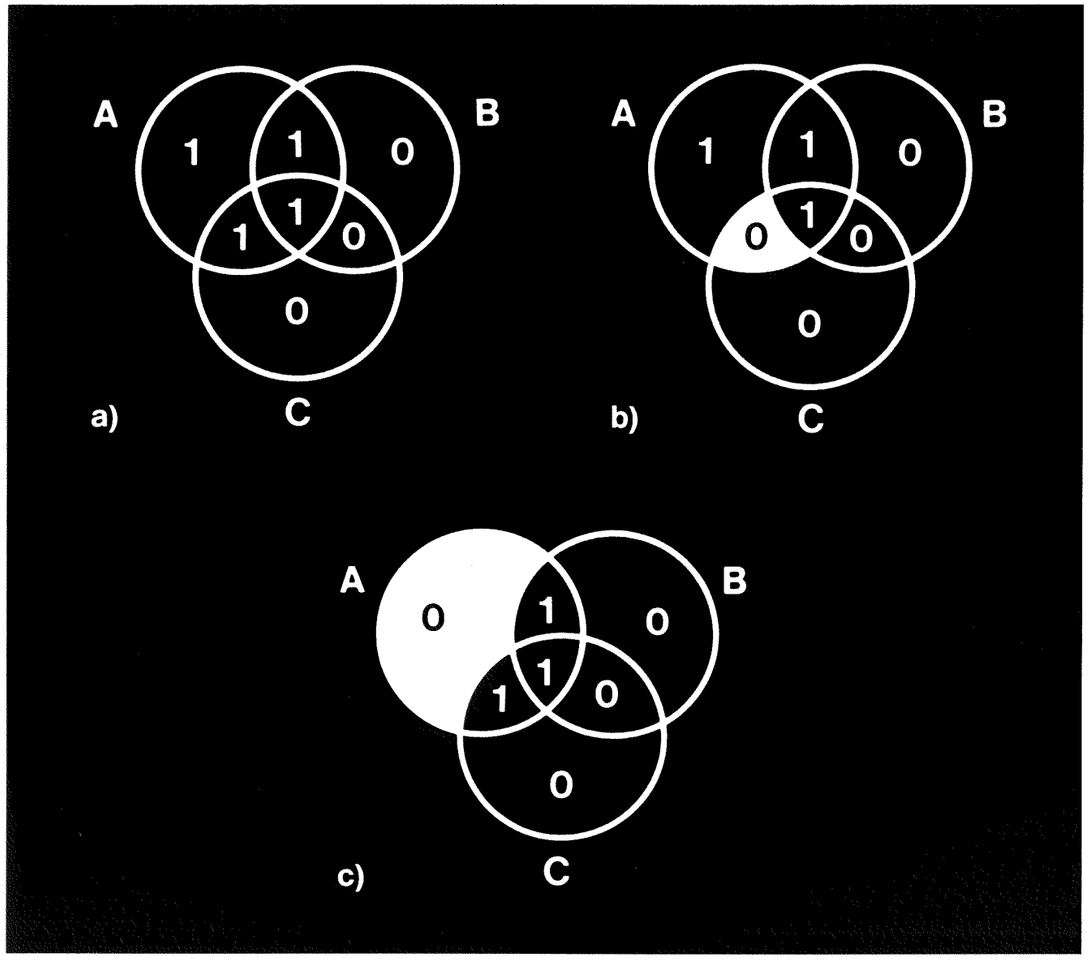
Il *numero* di bit aggiuntivi dipende dalla lunghezza della parola da memorizzare: per proteggere da 2 a 4 bit se ne aggiungono 3, per proteggerne da 5 a 11 se ne aggiungono 4, e così via. In questo modo il numero di bit aggiuntivi cresce, al crescere della parola da memorizzare, secondo una funzione a gradino. Il *valore* dei bit aggiuntivi viene determinato sommando (modulo 2) i valori dei bit in posizioni fissate della parola da memorizzare. Questi bit vengono concatenati alla parola e memorizzati assieme ad essa. In fase di lettura il controllo di questi bit permette di identificare l'eventuale errore.

Questo procedimento verrà illustrato con un semplice esempio. Si supponga di voler memorizzare quattro bit di informazione (ad es. 1110). La memoria è rappresentata da un diagramma (di Venn) formato da tre cerchi che si intersecano come è illustrato dalla figura 2.

I tre cerchi individuano un totale di sette regioni. I quattro bit di informazione vengono assegnati alle quattro regioni più interne del diagramma. Alle tre regioni restanti si attribuiscono i bit ottenuti sommando (modulo 2) i bit contenuti in ciascun cerchio. Al termine del procedimento, ciascuna delle sette regioni contiene un bit (cfr. fig. 2.a). Complessivamente i sette bit costituiscono la parola da memorizzare.

Si supponga ora che un errore logico modifichi uno dei bit della parola

Fig. 2 - Comportamento del codice di Hamming in presenza di 1 errore.



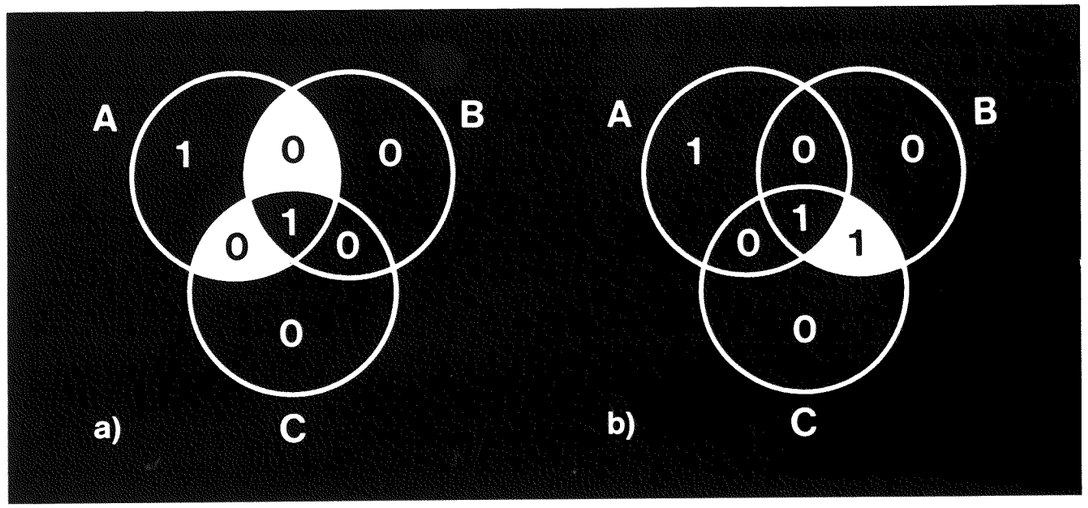
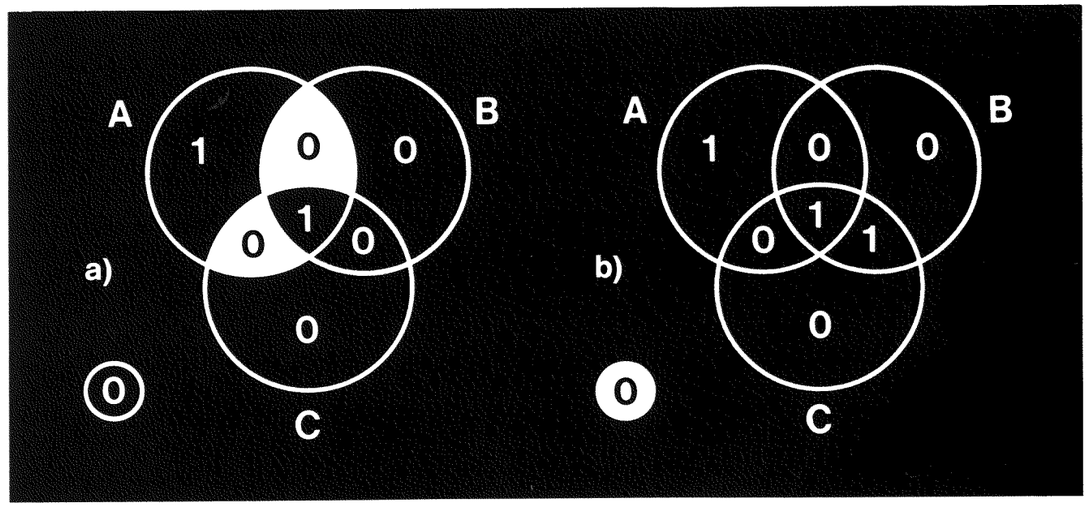


Fig. 3 - Comportamento del codice di Hamming in presenza di 2 errori.

Fig. 4 - Comportamento del codice di Hamming esteso in presenza di 2 errori.



(evidenziato dal diverso sfondo in fig. 2.b). Tale errore viene subito individuato poiché è sufficiente leggere i bit di controllo. Facendo riferimento alla figura, i bit di controllo informano che c'è qualcosa che non va nei cerchi A e C ma non in B. Dato che solo una delle sette regioni è contenuta sia in A che in C ma non in B, è facile individuare l'errore e correggerlo.

Più formalmente, in fase di lettura si verifica che i bit di controllo contengano realmente i valori ottenuti sommando bit in determinate posizioni della parola da controllare. I bit di controllo sbagliati permettono di individuare quale bit è errato.

È immediato anche rivelare se l'errore si manifesta in uno dei bit di controllo. Nella figura 2.c uno dei bit di controllo è stato invertito. Siccome si nota che l'errore riguarda solo il cerchio A, si modifica l'unico bit contenuto solo in quel cerchio (il bit di controllo errato).

Il procedimento descritto ha però un limite: ha successo solo quando la parola di codice ha subito un solo errore. Se si presentano duo o più errori il procedimento fallisce perché si tenta di modificare un bit che di solito è corretto. Si osservi la figura 3.a in cui sono evidenziati i due bit errati. Il meccanismo di correzione nota che qualcosa non va nei cerchi B e C, ma non nel cerchio A; quindi corregge un bit esatto (cfr. fig. 3.b), peggiorando ulteriormente la situazione.

3.2. Il Codice di Hamming Esteso

È possibile migliorare il codice di Hamming appena illustrato rendendolo capace di riconoscere il verificarsi di due errori.

Si introduce, all'esterno dei tre cerchi, un ulteriore bit di parità in modo che il numero totale di bit uguali a 1 nel diagramma sia pari. Un errore singolo viene riconosciuto e corretto grazie al meccanismo di correzione appena illustrato. Se invece se ne presentano 2 (cfr. fig. 4.a), viene corretto (dopo l'esame dei singoli cerchi) un bit esatto. Questa situazione di errore però viene rilevata dal bit di parità (cfr. fig. 4.b). Il codice è dunque in grado di riconoscere che si sono verificati due errori, ma non sa in quali posizioni si sono presentati e quindi non è in grado di porvi rimedio.

4. Occupazione di spazio e protezione dagli errori.

Si conclude riportando alcune considerazioni sull'occupazione aggiuntiva di memoria e sull'affidabilità di una memoria centrale di 1 Mbyte protetta del codice di Hamming esteso.

Si valuta dapprima lo *spazio* di memoria dedicato ai bit di controllo. Una parola di memoria di 32 bit, quella di solito usata, richiede 7 bit di controllo. In totale è necessario memorizzare 39 bit. Il rapporto tra la lunghezza della stringa di controllo e la lunghezza della parola (7/39) fornisce la percentuale di memoria dedicata ai bit di controllo, che è circa il 18%. Tale percentuale potrebbe sembrare elevata, ma è più che giustificata dall'affidabilità che ne deriva.

L'*affidabilità* delle memorie è collegata al concetto statistico di tempo medio prima di un guasto, il tempo che deve mediamente passare prima che una memoria perfettamente funzionante si guasti. Come accennato nell'introduzione, esperimenti di laboratorio stimano il tempo medio prima di un guasto in una singola cella di memoria a semiconduttore pari a circa 1.000.000 di anni, e quindi lunghissimo. Per l'intera memoria (senza codice di Hamming) il tempo medio scende a 43 giorni circa (perché si divide 1.000.000 di anni per il numero di bit della memoria) e si riduce a 35 giorni circa se la memoria è protetta dal codice di Hamming esteso.

In quest'ultimo caso però ogni parola di memoria può tollerare un errore in uno qualsiasi dei suoi bit: un errore in grado di influenzare il corretto funzionamento della memoria avviene solo quando in una parola si presentano almeno due bit errati. Calcoli statistici rivelano però che prima che la probabilità del verificarsi di questo evento (due errori in una stessa parola) superi il 50% devono mediamente passare 63 anni.

Considerazioni analoghe si possono fare anche nel caso di memorie di massa o qualora si utilizzino parole di memoria di lunghezza differente.

È evidente a questo punto la convenienza e la necessità dei codici correttori e rivelatori di Hamming per garantire l'affidabilità delle memorie centrali e di massa degli elaboratori.

5. Bibliografia

[1] Franco Filippazzi, *"Tecnologia dell'Elaboratore Elettronico"*, Franco Angeli Editore, 1983

[2] Giuseppe Longo, *"Teoria dell'Informazione"*, Boringhieri, 1980

[3] Robert J. McEliece, *"L'Affidabilità delle Memorie dei Calcolatori"*, Le Scienze, numero 199, marzo 1985.

Il computer nel portafoglio

EUGENIO CASUCCI

Honeywell Information Systems Italia
Milano

1 - Introduzione

Le tessere di materiale plastico, che siamo abituati a maneggiare ed utilizzare quasi quotidianamente, hanno subito una notevole evoluzione durante gli ultimi anni e sono, oramai, giunte ad una svolta decisiva: da supporti inerti sui quali memorizzare un numero molto limitato di informazioni si stanno infatti trasformando in veri e propri computer da tasca.

Il termine da tasca, in questo caso, non sta ad indicare una dimensione molto contenuta ma la reale possibilità di inserire la capacità elaborativa di un microprocessore entro una carta plastica meglio nota come carta di credito.

L'evoluzione che ha portato le tessere plastiche a questo sofisticato livello di funzionalità ed integrazione è passata, negli anni scorsi, attraverso numerose fasi caratterizzate dalle tessere con specifiche uniformate secondo gli standards internazionali, alle tessere con caratteri a rilievo, dalle tessere con fotografia incorporata a quelle a striscia magnetica.

Queste ultime, tra l'altro, sono le uniche che hanno permesso il primo vero passo verso uno strumento di identificazione personale riconoscibile da un computer; non riteniamo infatti molto significative le esperienze con tessere dotate di codice a barre alla luce del fatto che la loro diffusione è stata limitata ad alcuni settori di utilizzo molto specifici.

Le carte plastiche a striscia magnetica possono includere molteplici tipi di "piste" quali, ad esempio, le ISO e le TRANSAC.

Le piste ISO sono quelle poste nella parte alta del retro della carta e vengono normalmente utilizzate nelle applicazioni bancarie (es. BANCOMAT), parabancarie (es. VISA/EUROCARD) o per la rilevazione e la gestione delle presenze (es. applicazioni integrate tipo CARP).

Le piste TRANSAC sono quelle poste nella parte centrale del retro delle carte ed hanno un utilizzo più limitato delle piste ISO; in Italia, in particolare, un interessante uso di piste collocate nella posizione standard TRANSAC si ha nell'applicazione VIACARD per il pagamento dei pedaggi autostradali.

Le piste ISO, da un punto di vista più tecnico, hanno una capacità variabile che viene determinata dalla densità di registrazione delle informazioni; a titolo esemplificativo le capacità massime delle piste ISO sono le seguenti:

- traccia 1: caratteri 79
- traccia 2: caratteri 40
- traccia 3: caratteri 107

Anche le caratteristiche fisiche di queste carte sono fissate secondo gli standard ISO e riguardano, ad esempio, le dimensioni, il numero ed il tipo di torsioni a cui la carta deve essere soggetta senza rompersi o le temperature, minima e massima, entro cui le informazioni memorizzate sulla striscia magnetica devono mantenersi inalterate ed essere utilizzabili.

Un caso anomalo, rispetto al resto del mondo, è la realtà giapponese che prevede la striscia magnetica nella parte anteriore della carta anziché in quella posteriore.

Alle carte con striscia magnetica si sono oggi aggiunte carte con funzionalità più elevate soprattutto dal punto di vista della sicurezza, più che mai sentita dalle organizzazioni bancarie e parabancarie internazionali.

Il primo passo in questa direzione è stato quello di rendere più difficilmente falsificabile la base della carta; nella plastica, infatti, sono oggi correntemente inserite zone con inchiostro sensibile ai soli raggi ultravioletti ed aree riservate all'inserimento di ologrammi tridimensionali, come avviene per le carte VISA ed EUROCARD destinate ad un utilizzo internazionale.

Il secondo passo può essere l'adozione di strisce magnetiche "filigranate", che richiedono però una speciale testina di lettura ed un trattamento particolare delle informazioni aggiuntive, non presenti sulle piste "ISO" di tipo standard.

Prima di passare all'esame delle carte a microprocessore, desideriamo qui accennare anche all'esistenza delle carte al laser che, ad oggi, non

ci pare abbiano ancora raggiunto un sufficiente livello di maturità; delicatezza del supporto e scarsità di apparati per la lettura e la scrittura, producibili in serie ed a costi contenuti, sono gli ostacoli principali alla diffusione di questa tecnologia annunciata nel 1981 e non ancora applicata da nessuna organizzazione bancaria.

2 - La "carta intelligente" CP8

Le carte che includono un vero e proprio circuito integrato hanno invece superato la fase dell'infanzia e sono ormai entrate in quella della maturità che ne consente l'utilizzo su larga scala.

Si può affermare che le applicazioni effettive di questa tecnologia sono ormai una realtà all'estero e che anche in Italia c'è un grande interesse per queto tipo di prodotto.

Le carte che includono un "chip" di silicio al loro interno si devono però distinguere in due grandi categorie: quelle che sono dotate di sola memoria e quelle che includono anche un microprocessore per il trattamento delle informazioni.

Le carte dotate di sola memoria costituiscono un passo intermedio tra la striscia magnetica e la carta intelligente, anche se non sfruttano appieno la potenzialità di questa innovazione tecnologica.

Le carte a sola memoria vengono oggi utilizzate correntemente in Francia per il pagamento delle comunicazioni telefoniche delle cabine dotate di "publiphone a carte"; la tecnica di impiego è simile a quella in atto presso le nostre cabine con telefoni a carte magnetiche ed il grosso vantaggio di questa soluzione è la compatibilita dello strumento con le schede, più potenti, dotate di microprocessore.

Con queste ultime, ad esempio, è possibile telefonare facendosi addebitare l'importo della comunicazione sul conto di casa, o dell'ufficio, oppure eseguire il caricamento di quaranta scatti da utilizzare poi con scalamento automatico; queste carte, ovviamente, possono poi essere ricaricate con altri quaranta scatti per volta fino alla saturazione totale della memoria disponibile.

Entriamo ora un poco di più nella storia e negli aspetti costruttivi delle carte a microprocessore, per meglio comprenderne caratteristiche e potenzialità.

L'idea di una carta includente un "chip" con funzioni di trasferimento dati, elaborazione e memorizzazione si deve a R. Moreno, un ingegnere francese, che negli anni '70 raggiunse un accordo con la società CII-Honeywell-BULL perché questa ultima ne curasse la trasformazione in prodotto vero e proprio.

La tecnologia non si limita alla sola carta, coperta da circa quaranta brevetti, ma si estende anche ai dispositivi di lettura e scrittura delle carte stesse ed alla costruzione logica delle funzionalità delle carte, detta personalizzazione.

I passi per arrivare alle carte attualmente in circolazione sono stati piuttosto lunghi ed uno dei più difficili da compiere è stato quello per arrivare alla produzione di tessere che, pur rispettando appieno le norme ISO per le carte a pista magnetica, includessero un "chip" perfettamente funzionante.

Per fare questo si sono dovuti realizzare dei "wafer" con uno spessore inferiore a quello utilizzato per produrre dei "chip" standard; si è dovuto, infatti, ridurre lo spessore da 0,015 a 0,011 pollici affinché una carta secondo standard ISO potesse alloggiare un "chip".

Un altro grosso problema da risolvere è stato quello delle dimensioni del "chip", che si sono dovute contenere al massimo onde evitare i problemi legati alle sollecitazioni meccaniche cui il "chip" può venire sottoposto durante l'utilizzo.

La scelta dei progettisti, di conseguenza, si è orientata verso tecniche

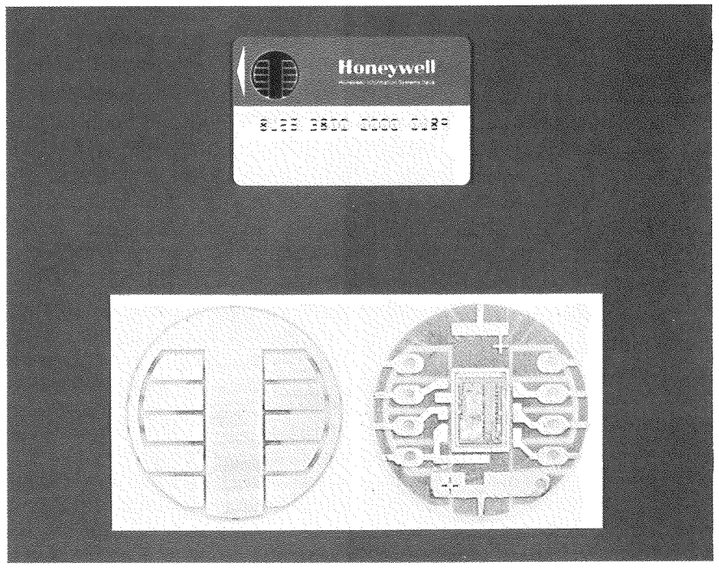
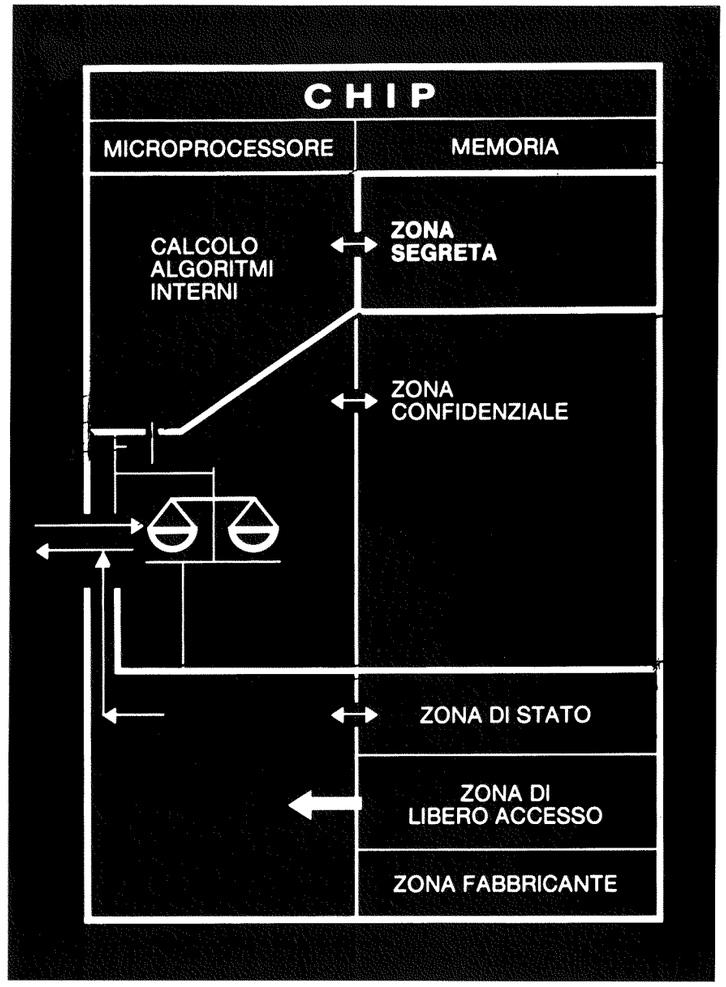


Fig. 1 - *In alto:* una "carta intelligente", tipo CP8.
Sotto: a sinistra, ingrandimento dei contatti della carta; a destra, ingrandimento del "chip" su cui è realizzato il microprocessore.

Fig. 2 - Schema funzionale del chip di un carta intelligente CP8.



microelettroniche che consentissero un impaccamento elevatissimo, rispettando però un requisito fondamentale, e cioè che fossero compatibili con un costo di produzione molto basso.

Tra i problemi tecnici incontrati nel passaggio dalla fase di idea a quella di prodotto, si deve ricordare la ricerca di soluzioni circuitali che consumassero il minimo possibile di energia elettrica, pur presentando un sufficiente velocitá ed un elevata insensibilità ai disturbi elettromagnetici.

Basta pensare ai disturbi provocati dai dispositivi elettronici ed elettromeccanici delle pompe di benzina e delle autovetture che si fermano per fare rifornimento, per comprendere quale livello di insensibilità ai rumori debbano avere delle carte utilizzate per effettuare il pagamento del carburante.

Il risultato finale dello sviluppo è stata la "carta intelligente CP8", realizzata dalla CII-Honeywell Bull, la prima del genere al mondo.
Il "chip" contenuto nella carta CP8 corrisponde ai molteplici requisiti posti, e può essere interessante fornire qualche dettaglio in merito, in quanto questo tipo di carta costituisce, ad oggi, l'unico standard di fatto utilizzato in molti paesi, primo fra tutti la Francia.

Queste carte hanno, come abbiamo già detto, le stesse dimensioni delle carte utilizzate negli apparati per la distribuzione di contante (0,76x54x85 mm), accettano la presenza di zone di rilievo, di ologrammi, di piste magnetiche ISO e TRANSAC ma, presentano delle condizioni di esercizio notevolmente più ampie e sofisticate delle normali carte a pista.

A titolo di esempio, si può ricordare come il contenuto della memoria del "chip" non venga alterato da campi magnetici fino al valore di 10.000 gauss (dieci volte superiore a quello fissato dagli standard per le carte magnetiche) e come possa resistere a radiazioni fino a 2000 rads (duemila volte superiori a quelle utilizzate nei controlli aeroportuali di sicurezza).

La collocazione del "chip", e dei relativi contatti, nella parte superiore

sinistra della faccia anteriore della carta, permette una elevata resistenza alle torsioni poiché negli angoli le sollecitazioni sono inferiori a quelle che si hanno nelle altre zone della carta.

Le soluzioni giapponesi, che si stanno affacciando ora sul mercato, prevedono invece la collocazione dei contatti nella parte centrale della carta, per compatibilità con la disposizione delle piste magnetiche nelle carte locali; ma ciò non ne impedirà, in futuro, l'utilizzo anche nei paesi occidentali perché la disponibilità di testine di lettura polivalenti permetterà di leggere tutti i tipi di carte senza problemi.

La posizione dei contatti è comunque un elemento da non sottovalutare e lo stesso si dica per il cosiddetto protocollo di colloquio della carta.

La norma ISO/DP 7816, in corso di approvazione, stabilisce proprio le regole che permetteranno ai costruttori di carte a microprocessore l'intercambiabilità del loro prodotto con gli altri esistenti.

Il "chip" della SMART CARD CP8 ha bisogno, per funzionare, di sei soli punti di contatto, che vengono selezionati durante la fase di produzione, invalidando tutti gli altri contatti.

I sei contatti rimasti attivi sono in grado di garantire le necessarie ridondanze sia per quanto riguarda i punti di alimentazione che quelli di ingresso ed uscita dei dati.

3 - La struttura del chip

Il chip è costituito, sostanzialmente, da quattro blocchi funzionali: una unità di governo, che gestice il flusso dei dati in ingresso e in uscita dalla carta, e tre memorie diverse, sia come compiti che come tecnologia.

Esse sono: (a) Una memoria RAM di 32 parole, con funzione di "blocco per appunti", per archiviare temporaneamente i risultati di calcoli intermedi; questa memoria è "volatile", cioè il contenuto si perde quando non c'è sorgente di energia elettrica. (b) Una memoria ROM (Read-Only Memory), non volatile di 1,6 K Byte, che non può essere modificata; questa memoria è programmata dal costruttore e contiene programmi di gestione dell'intero chip. (c) Una memoria PROM, ossia una ROM programmabile, di 8 K bit; questa memoria è non volatile (e quindi i dati non vanno persi quando non c'è tensione eletttrica sul chip), però può essere scritta durante l'uso della carta.

La memoria di tipo PROM ha un ruolo fondamentale nel sistema. Essa è divisa in zone, con funzioni diverse, che vengono definite durante il processo di "personalizzazione" del chip, di cui si parlerà più avanti. Esaminiamo le caratteristiche e i compiti di queste varie zone.

La zona del "fabbricante" è quell'area di memoria che viene scritta nel momento della nascita del chip; essa contiene il numero di serie del componente e lo rende, così, diverso da qualsiasi altro esistente

La zona di "libera lettura" viene scritta all'atto della personalizzazione del chip e contiene dati accessibili da parte di chiunque sia dotato di un apparato di lettura con un programma compatibile con quello residente sulla carta; queste informazioni, di solito, possono coincidere con quelle stampigliate a rilievo sulla carta nella zona a questo scopo destinata dagli standard ISO.

L'area di "controllo degli accessi" contiene le informazioni che indicano quale è lo stato della carta; in altre parole questa zona riporta le informazioni che permettono, o meno, la lettura della carta stessa.

La carta, infatti, è di tipo attivo e la fase di verifica del codice segreto (algoritmo e programma di controllo del PIN, o "Personal Identification Number") si svolge all'interno del chip stesso, con la possibilità di blocco dopo tre errori successivi; questa funzionalità mette al riparo da frodi nelle quali la ricerca del codice segreto venga effettuata con sistemi elettronici.
L'area di "transazione", è la parte di memoria che viene, normalmente, destinata ad accogliere le informazioni per l'applicazione utente alla quale essa è stata destinata.

La zona "segreta", infine, è la parte della memoria che garantisce l'inalterabilità del contenuto più riservato e l'assoluta induplicabilità della carta; in questa zona, durante la fase di personalizzazione, vengono inseriti tutti i codici utili per la base più segreta dell'applicazione, come gli algoritmi di decifrazione del codice segreto o il PIN, protetti dal fatto che il programma inserito nel microprocessore non ne consente la successiva estrazione.

Questa zona, infatti, viene utilizzata dal microprocessore per l'accesso ai dati segreti; questi ultimi però, non vengono mai "estratti" ma solo "confrontati".

Per quanto riguarda il trattamento della carta CP8, essa viene letta e scritta secondo standard predefiniti attraverso una testina multifunzionale che è inserita in un apposito connettore; questo viene collegato ad una piastra che contiene i componenti atti alla lettura ed alla scrittura della memoria della carta, sempre tramite il microprocessore inserito nel chip.

L'insieme così ottenuto si può a sua volta inserire in apparati diversi, quali terminali per collegamenti standard, ad esempio con personal computer, oppure terminali per il pagamento elettronico (EFT: Electronic Fund Transfer).

4 - Personalizzazione della carta

Il processo di fabbricazione delle carte a microprocessore è costituito da due fasi: la prima è la vera e propria realizzazione fisica della carta e la seconda è la personalizzazione del chip in funzione dell'applicazione desiderata.

La costruzione della carta è piuttosto semplice: consiste nell'inserire la "pastiglia" (ossia il chip saldato alla contattiera) in un apposito allogiamento scavato nella tessera di plastica.

La personalizzazione richiede diverse operazioni.

La prima parte della personalizzazione è costituita dalla scrittura, nel microprocessore, del programma di trattamento dei dati; questo programma prevede la gestione delle zone della memoria in funzione del tipo di applicazione prevista; il programma per la gestione delle carte bancarie, ad esempio, è differente

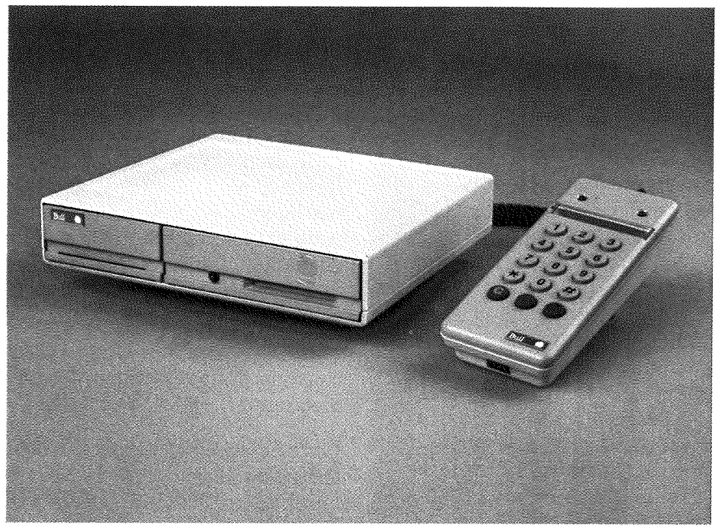


Fig. 3 - Lettore di carte intelligenti.

da quello utilizzato nelle carte per uso sanitario o per la patente elettronica.

Il programma, ovviamente, può essere redatto in funzione di specifiche esigenze dell'utente anche se, di norma, si ricorre ad una serie di soluzioni standard che consentono il trattamento della memoria nel modo consono alla maggior parte delle esigenze.

Nella seconda fase della personalizzazione la memoria viene predisposta per la ricezione dei dati applicativi; ciò significa delimitare le zone, predisporne il tracciato come se si trattasse di un comune archivio su disco, etc..
Questa fase viene chiamata "mappatura" e precede quella di inserimento dei dati da congelare nelle zone "segrete" e di "libera lettura". A maggior garanzia dell'utente, il sistema di personalizzazione del chip può ricevere i dati chiave sia dal programma che da una carta, precedentemente predisposta, da inserirsi all'interno della macchina nel momento stesso della personalizzazione di ogni gruppo, o lotto, di carte.

In questo modo, infatti, una banca può creare una carta "madre" contenente i codici da inserire in tutte le carte che da essa verranno generate senza che alcuna persona, tra tutte quelle che operano nell'ambiente delle carte di pagamento, possa conoscere le chiavi degli algoritmi di riconoscimento dei codici segreti o del P.I.N..

Per alcune applicazioni, invece, il processo di personalizzazione si può arrestare al momento nel quale vengono "mappate" le aree di memoria in quanto i codici identificativi del titolare non sono ancora noti.

In questo caso si potranno inserire dei codici chiave, segreti, validi per alcuni gruppi di carte lasciando ad una fase successiva, detta di "inizializzazione", il compito di assegnare la carta al titolare della stessa.

Una applicazione di questo tipo prevede carte da consegnare a persone non note fino al momento della distribuzione della carta stessa come ad esempio, in una USL, alle donne in stato di gravidanza. In questo caso ogni USL potrebbe avere un lotto di carte prepersonalizzate con la codifica relativa all'ente emittente ma la scrittura delle informazioni sulla gestante verrebbe fatta sullo stesso sistema con il quale verranno in seguito aggiornate le carte, direttamente presso la struttura sanitaria.

5 - Le apparecchiature per il trattamento

Per il processo di personalizzazione delle carte CP8 sono state messe a punto delle apparecchiature, dette ALEC, integrate in un sistema modulare basato su di un sistema DPS6 al quale possono venire collegate, oltre alle normali periferiche, delle stampanti per la spedizione di lettere associate alle carte (per esempio codice segreto), delle apparecchiature per la magnetizzazione della striscia, se presente, e per la punzonatura dei dati a rilievo, se richiesta.

Questo sistema è, ovviamente, basato su di un software specializzato che permette la gestione contemporanea ed integrata di tutte le apparecchiature sopra descritte.

Il sistema, grazie alla sua modularità, può collegare fino a cinque ALEC contemporanei che permettono di quintiplicare la produttività dello stesso. I dati in ingresso, ovviamente, pervengono dal programma parametrizzato, per quanto riguarda la mappatura della memoria, e da supporto magnetico (nastro), più la carta "madre", per quanto concerne le informazioni sull'ente emittente ed il portatore.
Dopo l'emissione della carta diviene necessario poter disporre di una serie di apparati per la lettura, la scrittura e per l'aggiornamento delle carte.
A questo proposito sono disponibili diversi livelli di hardware che consentono di mantenere la massima flessibilità e modularità del sistema.
Il primo livello di apparati è costituito dai "connettori", dispositivi atti al

trattamento fisico della carta, che devono essere collegati, ovviamente, ad una "logica" di trattamento.

Un complesso elettronico, detto "coupleur", costituisce il secondo livello di integrazione, e mette a disposizione dell'utenza una piastra completa di connettore e logica di trattamento; questo apparato lavora secondo un ben preciso protocollo di colloquio e può essere facilmente integrato in terminali, dispositivi od altre apparecchiature.

Il terzo livello di integrazione è costituito da apparati che, essendo dotati di logica di colloquio con sistemi standard come un personal computer, possono fungere da lettori/scrittori dell carte CP8.

È questo il caso dei lettori che consentono l'utilizzo in ambiente MSDOS, in collegamento con una normale piastra asincrona dotata di porta RS232/V24.

Apparati analoghi, ma con logica diversa, vengono collegati ai terminali "TELETEL", che in Francia ed in USA sono utilizzati per le applicazioni videotex di tipo standard; in questo caso i lettori hanno al loro interno una piccola cartuccia che contiene un programma applicativo pari a 24 KBytes.

Grazie a questo programma un semplice schermo di consultazione videotex si trasforma in una vera e propria stazione di trattamento delle carte a microprocessore CP8. Questo hardware, però, corrisponde agli standard di telecomunicazioe videotex francesi (teletel) e pertanto non è utilizzabile, in Italia, se non in modo off-line.

Le ipotesi di applicazioni, però, fanno privilegiare utilizzi nei quali sia presente almeno un minimo di capacità elaborativa e di memorizzazione locale. In altri termini la soluzione più promettente sembra essere proprio quella con un personal computer che, localmente, svolge le funzioni di trattamento e gestione della carta, memorizzando le transazioni e consultando archivi di dati per trasmettere poi il tutto, in differita, al computer centrale.

Altre applicazioni prevedono invece l'integrazione logica del lettore nella tastiera, onde permettere la gestione di un accesso sicuro alle banche dati centralizzate.

Ulteriori tipi di apparati per il trattamento delle carte CP8 sono i terminali per il pagamento elettronico (EFT; POS), i terminali per la gestione delle timbrature e delle aperture di varchi, i telefoni pubblici a carta, questi ultimi diffusi solamente in Francia.

6 - Applicazioni

Esiste un amplissimo ventaglio di possibili applicazioni delle carte a microprocessore. A titolo di esempio, forniamo qui di seguito un elenco di alcune delle applicazioni della carta CP8 fatte in Francia.

Carta bancaria CB. Questa carta possiede tutte le funzionalità della carte di credito a pista magnetica (prelievo di contante, pagamento nei negozi) e offre in più nuovi servizi quali il telepagamento, la banca a domicilio, l'utilizzo dei telefoni pubblici attrezzati per tale mezzo.

La distribuzione di queste carte è iniziata da quest'anno e si prevede la consegna di oltre 16 milioni di esemplari entro il 1988.

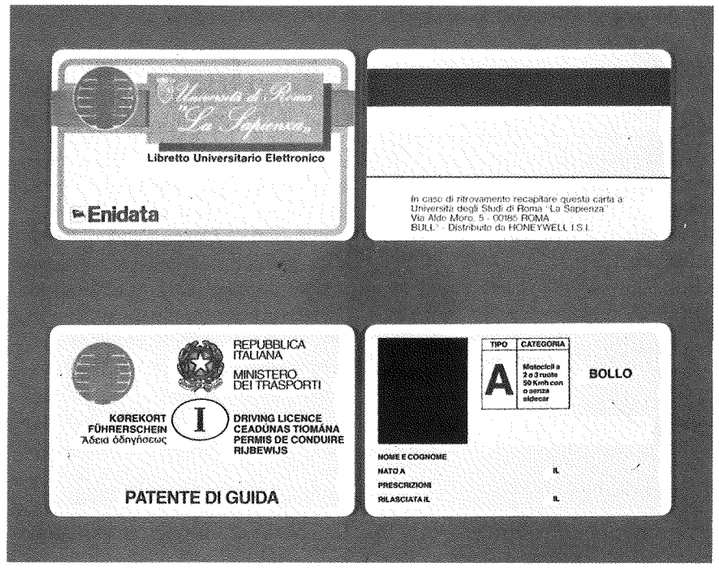
Carta Publiphone. Questa carta della Direzione Generale delle Telecomunicazioni Francesi (DGT) permette all'utente di telefonare da una cabina attrezzata. Si hanno due formati: carta prepagata o Telecarta senza codice confidenziale; carta di credito TELECOM, con codice confidenziale, che addebita le telefonate direttamete sulla bolletta telefonica dell'utente.

Carta SANEF. Consente il pagamento automatico del pedaggio sulla autostrada della SANEF (Società Autostrade Nord-Est della Francia).

Carta sanitaria di Blois. Costituisce la scheda medica portatile per i bambini fino a due anni (vaccinazioni, ecc.) per le gestanti (visite, analisi, ecc.) e per le persone oltre i 65 anni (contro indicazioni/farmacologi-

Fig. 4 - Un ventaglio di carte a microprocessore per gli usi più svariati.

Fig. 5 - Due esempi di applicazioni: libretto universitario e patente di guida.



che in caso di urgenza). La scrittura e la lettura delle carte è effettuata sotto il controllo della carta di abilitazione del medico.

Carta della Universtà Paris 7. Sostituisce il libretto universitario; contiene in particolare: le informazioni anagrafiche e amministrative, i corsi seguiti, gli esami sostenuti.

Ticket per trasporti urbani. Viene impiegata come mezzo per usare i trasporti pubblici di Blois.

Carta DDASS. Questa carta permette agli assistenti sociali del centro di assistenza di Bordeaux di accedere ai dossier degli assistiti, usando dei terminali Minitel.

L'elenco che abbiamo sopra riportato non è completo e serve a dare una idea della varietà delle applicazioni possibili con questo mezzo. La carta CP8 ha ormai una diffusione significativa non solo in Francia, dove è nata, ed a cui si riferivano gli esempi precedenti; anche in Italia questa carta a microprocessore va prendendo piede, e a conferma di ciò, si possono citare le applicazioni seguenti.

TELLCARD. Carta di pagamento utilizzata durante i campionati mondiali di sci del 1985.

Patente elettronica. Il dispositivo è in grado di memorizzare informazioni diverse, anche visive (fotografia, bollo di rinnovo, dati anagrafici). La carta puo essere letta con un terminale portatile di piccole dimensioni e di basso costo. L'applicazione è stata presentata durante il salone della automobile di Torino 1985.

Libretto universitario. Questo libretto elettronico viene distribuito, in via sperimentale, agli studenti dell'Università "La Sapienza" di Roma, dal maggio 1986.

Sicurezza fisica. Una applicazione, in atto in ambito militare, si avvale della carta CP8 per garantire la sicurezza degli accessi ad aree riservate. Il sistema permette di risolvere, in modo integrato, anche problemi gestionali come il pagamento di mense e la rilevazione delle presenze.

Libretto sanitario. La carta CP8 è stata scelta per realizzare il "libretto sanitario" sperimentale per gli assistiti di una USL della regione Veneto.

Questo elenco non è ancora completo, perché diverse altre applicazioni sono fase di sviluppo o di studio.

7 - Conclusioni

La carta a microprocessore è una spettacolare realizzazione della tecnologia moderna, che ci consente di tenere un vero e proprio computer nel portafoglio.

Per quanto riguarda il futuro, esistono orientamenti verso chip con maggiori capacità di memoria, nonché verso carte dotate di visore a cristalli liquidi e con tastiera incorporata.

A prescindere da ulteriori progressi, la carta a microprocessore costituisce già adesso un prodotto di grande interesse pratico, con una grandissima gamma di applicazioni già realizzate o in fase di sviluppo.

L'auditor EDP una nuova figura professionale

ROSARIO MARASCO
Honeywell Information Systems Italia
Milano

1. Introduzione

L'elaborazione elettronica dei dati, legata all'incessante sviluppo dei computer e alle loro applicazioni nei più svariati settori ha fatto sì che l'informazione sia ormai una risorsa aziendale e che come tutte le risorse (umane, finanziarie, tecniche, ecc.) debba essere opportunamente protetta.

Il patrimonio di informazioni della azienda non può correre il rischio di essere dissipato da un cattivo utilizzo dei mezzi EDP (Electronic Data Processing).

Occorre pertanto fornire all'azienda garanzie di contenimento dei rischi a cui essa viene sottoposta allargando l'utilizzo degli strumenti informatici per il trattamento delle informazioni.

Questa garanzia di sicurezza EDP costituisce la ragion d'essere della funzione EDP auditing.

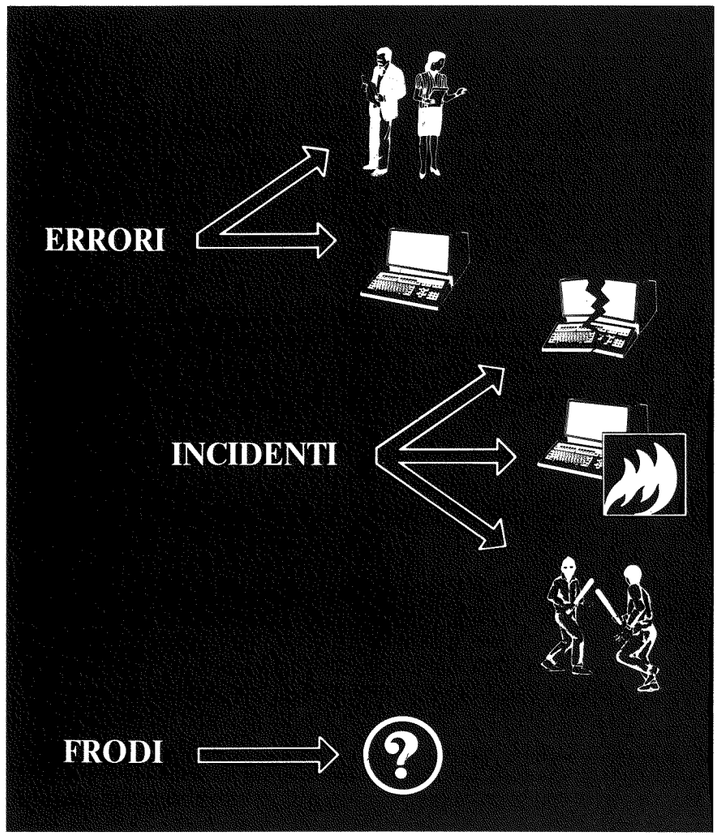


Fig. 1 - Eventi che possono danneggiare il patrimonio informatico di una organizzazione

La sicurezza dei dati in ambiente EDP è un problema che va affrontato e risolto ricorrendo ad iniziative, strumenti e tecniche di tipo diverso, coinvolgendo tutti coloro che operano nel sistema informativo in maniera diversificata a seconda che siano proprietari, utenti o gestori delle informazioni.

L'impegno più grosso è di questi ultimi, i quali, essendo i principali artefici del processo automatico devono, indipendentemente dall'aspetto tecnico ottemperare caso per caso alle precise indicazioni dei "proprietari" per l'applicazione selettiva delle misure di sicurezza in funzione della criticità e riservatezza dei dati.

Questa responsabilità non è delegabile e deve essere formalizzata nella documentazione di ogni applicazione che venga sviluppata.

Le misure di sicurezza da adottare non possono limitarsi ai packages da comprare e da installare (solo alcuni di essi hanno queste caratteristiche), ma sono da individuare ad hoc per ogni singolo ambiente costituito o da costituire, piccolo o grande, semplice o complesso.

2. Il "controllore" della sicurezza

L'EDP auditing non è una figura fiscale che effettua controlli sul comportamento del personale EDP, né tantomeno una funzione aziendale di pronto intervento in grado di far fronte a qualunque situazione di emergenza che possa verificarsi nell'ambiente EDP; essa è una funzione che svolge in modo sistematico prevalentemente attività di prevenzione contro eventi di tipo accidentale ed intenzionale che possono danneggiare il patrimonio informatico dell'azienda.

In sintesi gli eventi cui far fronte si possono raggruppare in 3 categorie (fig. 1):

errori (umani o delle macchine);

incidenti (alle macchine, agli ambienti umani);

frodi (attentato al patrimonio aziendale tramite le macchine).

L'ufficio cui è delegato il compito di controllore della sicurezza (auditing), deve assicurare che le risorse di una organizzazione siano protette in modo adeguato (come si può osservare sinteticamente nella fig. 2).

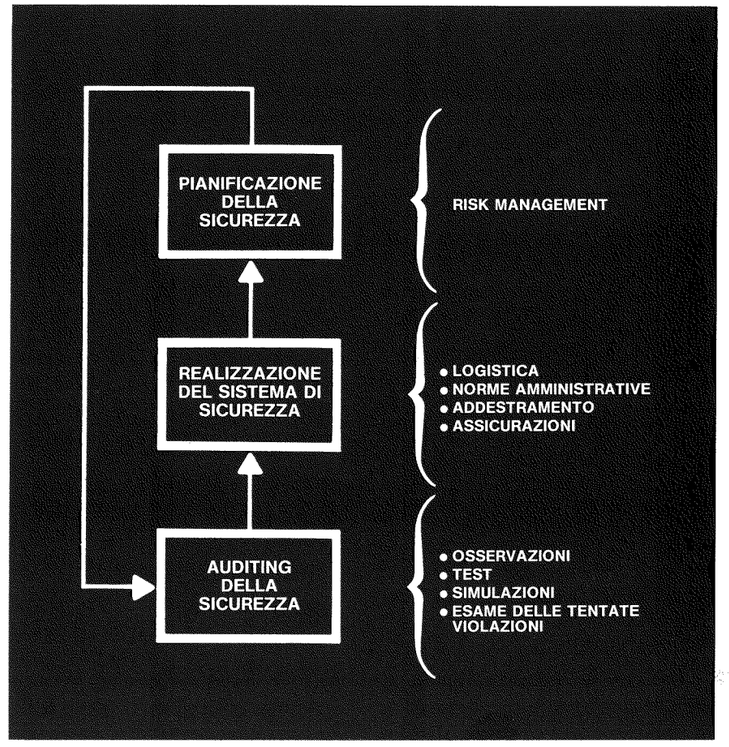
L'auditor non ha la funzione di scovare frodi od irregolarità, ma suo preciso compito è valutare e vigilare sull'applicazione delle procedure interne di controllo in conformità agli standard accettati; egli, se giudica insufficienti i controlli, ha il dovere di intervenire per suggerire le opportune modifiche.

3. Il campo d'azione dell'auditing

Una funzione elaborativa consta dei seguenti elementi:

- il sistema applicativo, costituito da:
  - procedure manuali per originare e trasmettere le transazioni di input al reparto DP (Data Processing)
  - programmi applicativi che controllano l'elaborazione delle transazioni, delle registrazioni e la preparazione dell'output

Fig. 2 - Il ruolo del controllore della sicurezza del sistema informatico (Auditor EDP)



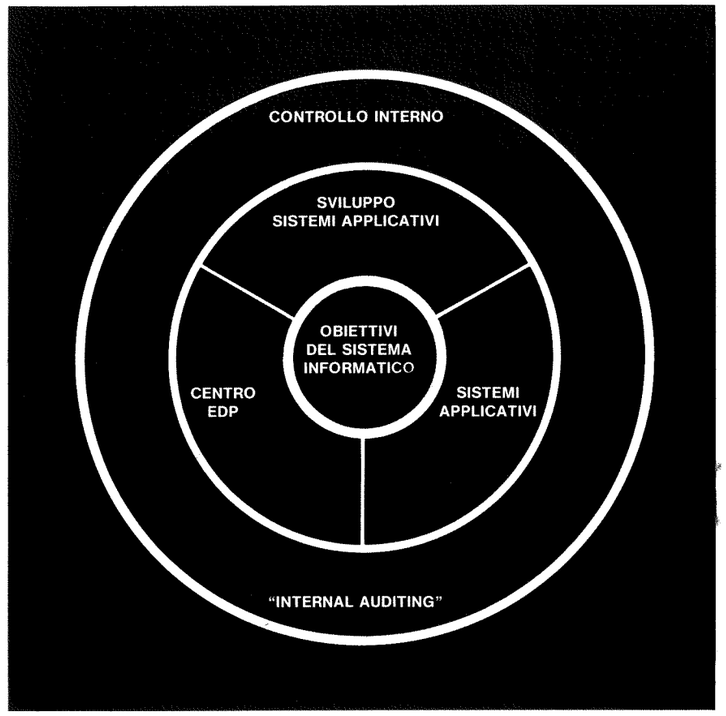


Fig. 3 - Il campo d'azione dell'auditing EDP.

- procedure che guidano il personale del centro EDP nell'uso di specifici programmi e dei relativi dati di input e di output

- le operazioni del centro, che riguardano:
  - l'ambiente
  - il personale
  - le macchine e le procedure generali che regolano le suddette operazioni, in contrapposizione alle procedure specifiche delle singole applicazioni
- lo sviluppo delle applicazioni, che investe:
  - il personale e le procedure inerenti il disegno
  - la stesura dei programmi
  - le prove e l'introduzione di procedure manuali e di programmi applicativi che rendono operativa l'applicazione; questi elementi comprendono anche la modifica ed il miglioramento dei programmi esistenti.

In figura 3 sono evidenziati questi elementi nella loro relazione con il controllo interno e con il mandato dell'EDP auditing; essi sono tra loro correlati, infatti vengono pianificati, organizzati e diretti per raggiungere gli obiettivi dei sistemi informativi.

Essi sono anche interdipendenti, ad esempio, lo sviluppo del sistema può essere determinato dalla disponibilità di maggior capacità elaborativa e di risorse specialistiche; così la capacità elaborativa può essere accresciuta con l'aggiunta di dispositivi richiesti da nuove esigenze emerse durante lo sviluppo.

Oggetto del controllo sono le risorse informatiche utilizzate per lo sviluppo, la manutenzione e la gestione del sistema informativo aziendale, cioè:

- il personale EDP e gli utilizzatori
- le risorse hardware
- il patrimonio software
- i dati memorizzati sugli archivi
- le norme, gli standard ed i metodi adottati nello sviluppo, manutenzione e gestione dei sistemi applicativi
- le misure di sicurezza logica e fisica
- le alternative di back-up per tutto l'hardware aziendale
- le procedure di salvataggio (recovery) del patrimonio informatico.

4. Le tecniche di controllo

Le tecniche che l'EDP auditor può applicare su un sistema informativo sono:

- domande di carattere generale per stabilire il livello di sicurezza
- liste di controllo che coprano i vari elementi:
  - documentazione
  - programmi
  - input/output
  - librerie, programmi e archivi dati
  - ecc.
- controlli locali
  (essi devono assicurare che gli archivi siano bilanciati e che le procedure di sicurezza e di integrità non vengano escluse)
- campionamento
  (scegliere alcune transazioni e seguirle, nelle loro elaborazioni, fintanto che le registrazioni dell'elaboratore lo consentono)
- esecuzione di transazioni errate
  (fornire al sistema transazioni non valide per controllare l'efficienza delle procedure di controllo)
- esame dei tentativi di violazione di sicurezza e di integrità
- simulazione di elaborazioni per verificare la correttezza di taluni programmi significativi (es. calcolo degli interessi nelle applicazioni depositi a risparmio e conti correnti).

La direzione aziendale, da parte sua, deve fare affidamento fondamentalmente sulla funzione dell'internal auditing.

In Italia tale figura oggi non è troppo conosciuta (anche se già presente in molte aziende) e con un ruolo spesso non ben delineato.

Se diamo uno sguardo all'estero, soprattutto agli USA, questa figura è già pienamente affermata e la convenienza del suo esistere e del suo operare è attestata da numerose imprese.

- EDP manager (documentazione utente).

L'obiettivo dei controlli EDP è di assicurare che i dati autorizzati siano completamente ed accuratamente elaborati; la forma ed i modi di applicazione dei controlli stessi sono invece suscettibili di modifiche.

Possiamo così articolare i controlli:

1) inseriti nel software applicativo
2) di integrità
3) inseriti nel software di base (sistema operativo)
4) inseriti nei sistemi di telecomunicazione (data communication)
5) espletati dagli utenti.

6. Le tipologie dei controlli

6.1. Controlli inseriti nel software applicativo

Il software applicativo è l'insieme di programmi atti a soddisfare le singole esigenze.

Tali programmi ed il loro uso possono essere saltuariamente sottoposti a controlli in modo da verificare la corrispondenza fra quanto viene fatto e quanto è effettivamente richiesto.

Ad esempio, le istruzioni dei programmi sono di interesse del revisore quando si riferiscono a:

operazioni di controllo
(es., controllo di sequenza automatico e la preparazione a stampa di un elenco di numeri mancanti, controlli di quadratura)

operazioni contabili
(es., il calcolo e la preparazione di fatture di vendita, la generazione di premi assicurativi per l'automatico rinnovo di polizze).

L'esistenza e l'effettiva utilizzazione di detti controlli può essere già verificata in fase di realizzazione del sistema applicativo e dalla sicurezza garantita nel loro uso.

Il sistema applicativo costituisce un patrimonio fondamentale per l'azienda e quindi le attività relative alla sua gestione e manutenzione debbono essere disciplinate al fine di assicurare adeguata protezione a tale patrimonio.

La documentazione più significativa in materia riguarda:

- gli standard e la normativa per il passaggio da test a produzione, e viceversa, dei programmi nuovi o modificati
- gli standard da seguire e la documentazione da produrre per prove, paralleli, accettazione
- la normativa per il back-up delle librerie
- le procedure di gestione del file di "log" delle modifiche.

Sulla base di tale documentazione l'auditor EDP dovrà verificare che solo modifiche autorizzate possano essere apportate ai programmi e che ciò avvenga ad opera di personale autorizzato; in particolare nessuno degli addetti alle unità operative del centro di elaborazione deve avere possibilità di effettuare modifiche.

Occorre verificare inoltre che la normale manutenzione dei programmi avvenga secondo metodi di lavoro che garantiscono la separazione delle funzioni tra ambienti di prova e di produzione e che sia facilmente possibile il recupero di vecchie versioni di programmi modificati.

L'attività dell'auditor EDP si esplicherà attraverso l'esame della rispondenza della normativa e delle procedure alle finalità di salvaguardia del patrimonio costituito dai programmi applicativi e nella verifica dell'osservanza costante ed attenta della disciplina prevista.

6.2 Controlli di integrità

L'integrità in un sistema informativo consiste nel permanere delle caratteristiche di correttezza dei dati sia durante la loro archiviazione sul supporto magnetico di memorizzazione, sia durante le manipolazioni che essi subiscono nelle varie fasi di elaborazione, sia durante il trasporto dei dati stessi sulle linee di comunicazione.

L'integrità dei dati implica la loro salvaguardia contro manomissioni volute o meno, guasti, perdita, ecc.

Per questo scopo deve risultare efficiente in un sistema informativo, la corretta gestione di tutti i supporti

5. I compiti dell'auditor

Compiti prioritari dell'auditor interno sono:

- verificare e valutare se i controlli contabili ed operativi sono adeguati e risultano applicati correttamente. In particolare l'auditore EDP deve:
- accertare la realizzazione delle direttive, delle procedure
- accertare che i beni aziendali siano salvaguardati da perdite di qualsiasi genere
- assicurarsi dell'affidabilità ed integrità del sistema informativo aziendale
- valutare la qualità di esecuzione dei vari compiti assegnati
- accertare il rispetto delle norme e disposizioni legislative (es. DPR n. 600/73)
- controllare (o meglio effettuare il "Bilancio informatico") mediante "CROSS REFERENCES" la corrispondenza tra i nomi formalizzati con:
  - librerie
  - programmi sorgenti
  - programmi oggetto
  - procedure
  - files e report

magnetici numerati univocamente contenenti:

- librerie sorgenti e oggetto
- procedure
- files e archivi elettronici

con tutte le versioni storiche precedenti (generazioni) come prescritto dal DPR 600/73, oltre che dalla corretta gestione aziendale. Una gestione bilanciata per "nomi" del contenuto e "numeri" del supporto.

I controlli di integrità vengono svolti per assicurare che:

- procedure appropriate siano inserite effettivamente nei programmi, sia in fase di realizzazione del sistema sia quando occorre apportare, successivamente, delle modifiche (controlli sullo sviluppo delle applicazioni)
- variazioni non autorizzate non siano apportate ai programmi in uso (controlli di sicurezza), né agli archivi
- le procedure inserite nei programmi siano regolarmente svolte (controlli sulla operatività o gestione operativa).

Ciò non è da confondere con il controllo di efficienza sull'impiego del computer, ma rientra nell'ambito dell'auditing.

6.2.1 Integrità degli archivi

Le tecniche di recovery, cioè quelle che consentono in un sistema di ripristinare uno stato corretto dopo il verificarsi di un'avaria, sono largamente usate negli attuali sistemi e stanno assumendo importanza sempre maggiore con l'estendersi delle applicazioni.

Un'avaria è un evento per cui un sistema non si comporta secondo le proprie specifiche.

Per poter far fronte ad essa il sistema può essere arricchito da ulteriori componenti e particolari algoritmi.

Le avarie possono essere causate da inconvenienti hardware (mancanza di corrente o avaria di un disco) o da errori software (errori nei programmi o dati non validi) ed anche da errori umani.

L'affidabilità di un sistema dipende essenzialmente dal modo con cui riesce ad ovviare al malfunzionamento, senza compromettere l'integrità dei dati che al momento dell'avaria stava trattando.

Un caso di recovery abbastanza singolare è il back-up di un elaborazione ovvero il rifacimento di un lavoro eseguito nel passato.

L'affidabilità di un sistema dipende dal modo con cui riesce a rielaborare retrocedendo nel tempo, grazie alla corretta gestione dei supporti magnetici storici, e a utilizzare tutti gli elementi di quel tempo (tabelle - cambi - valute - condizioni - tassi - codici ecc.).

6.2.2 Traccia di "audit"

Una traccia di audit registra su un archivio tutti gli eventi verificatisi nel processo elaborativo.

Tale archivio può, ad esempio, contenere informazioni sugli effetti delle operazioni eseguite, la data e l'ora in cui sono avvenute, il codice di identificazione dell'utente (o dei programmi utenti) interessato, nastri e dischi allocati e non utilizzati, ecc.

Questa tecnica può essere usata per scopi diversi come:

- ripristino da caduta di sistema se vengono tenute versioni di riserva degli archivi, la traccia di audit consente di effettuare normalmente gli aggiornamenti e ripristinare i loro stati al momento del guasto.
  Tuttavia ormai esistono inserite nel software di base tecniche più sofisticate e rapide di ripristino di sistemi.
  La traccia di audit può servire come ulteriore sicurezza in caso di avaria di questi sistemi.
- affidabilità, comunemente conosciuta come tecnica di check point/restart. Dopo un guasto di sistema che non abbia danneggiato la memoria secondaria (o di massa), è possibile effettuare dei controaggiornamenti atti a ripristinare il contenuto degli archivi in aggiornamento durante il processo elaborativo al momento dell'evento; gli stati ripristinati sono quelli esistenti in un preciso istante del processo elaborativo (checkpoint)
- controllo della correttezza del sistema: tramite l'analisi della traccia di audit è possibile verificare il rispetto di certe regole e politiche, espressioni di legge, accordi commerciali, ecc.

6.3. Controlli inseriti nel software di base

Il software di base comprende una serie di programmi non specifici di una particolare applicazione ma di supporto alle fasi di progettazione, elaborazione e controllo di qualsiasi tipo di applicazione.

L'elaborazione dei dati è resa possibile da un modulo software, costituito da un insieme di programmi, che prende nome di sistema operativo (SO).

Il sistema operativo funziona a due livelli di attività:

- livello di utenza
- livello di supervisione.

Il primo livello corrisponde alla condizione in cui tutti gli utenti possono accedere alle funzioni autorizzate ed in cui sono operative tutte le eventuali misure di sicurezza (quali password, accessi limitati, ecc.).

Il secondo consente di accedere ai segreti più nascosti del sistema.

Pertanto è indispensabile proteggere l'accesso a questo livello con grande rigore, considerandolo alla stregua di un ingresso di una cassaforte e di conseguenza dotandolo di password multiple da usarsi compiutamente con persone diverse.

A tale livello si può scavalcare qualsiasi parola d'ordine, si può leggere ogni file, si possono dare o togliere autorizzazioni, si possono alterare, creare o distruggere informazioni ed istruzioni già impartite.

Ad esempio, i programmi applicativi in uso possono essere protetti da variazioni non autorizzate, mediante l'uso di apposite istruzioni inserite nel software che gestisce le librerie; l'uso corretto degli archivi può essere assicurato da controlli sulle "labels" ed il corretto svolgersi delle elaborazioni può essere fatto dipendere da procedure inserite nel sistema operativo e nel software collegato.

L'attività volta ad acquisire, installare, modificare, mantenere, ottimizzare e controllare le risorse di base ossia il software di base è necessaria per la gestione operativa della elaborazione dei dati.

La documentazione più significativa da acquisire può essere la seguente:

- piani di installazione del software
- procedure di autorizzazione ed accettazione delle modifiche al software
- procedure di rilevazione dei malfunzionamenti e del livello di assistenza assicurato dai fornitori
- registro degli interventi correttivi e/o di ottimizzazione.

L'attività di revisione, effettuata sulla base della suddetta documentazione, è diretta a verificare l'esistenza di norme e procedure che guidino tutte le fasi di modifica o di aggiornamento del software di base al fine di evitare discontinuità di servizio, alterazioni o distruzioni del patrimonio informativo e disallineamenti; in particolare si dovranno esaminare le procedure di prova per accertare l'esistenza e la validità degli ambienti di test.

Si dovrà inoltre verificare che vengano effettuate adeguate rilevazioni sul livello di servizio e sulle "performance" e vengano eseguite le conseguenti attività di manutenzione ed ottimizzazione.

Particolare cura dovrà essere posta a che le persone addette all'attività in esame siano dotate, normalmente, di elevata conoscenza tecnica ed esperienza e debbano, necessariamente, godere di ampia possibilità di intervento.

Questo però non deve tradursi nella possibilità di eludere le misure di controllo, le quali debbono comunque essere in grado, quantomeno, di segnalare le "forzature" e di individuare il responsabile.

6.4. Controlli inseriti nei sistemi di telecomunicazione

L'organizzazione che ha adottato un sistema di comunicazione tende a soddisfare i seguenti obiettivi:

- assicurare che adeguati controlli siano applicati a supporto di ogni applicazione interessata alla trasmissione, sia che questa avvenga entro l'ambito dell'organizzazione, sia che si faccia uso di linee telefoniche
- verificare che i controlli in atto funzionino come richiesto
- individuare e soddisfare le mutate esigenze di controllo, derivanti da ogni tipo di cambiamento.

Le esigenze elaborative attuali richiedono il trattamento di un elevatissimo numero di informazioni, in tempi brevi, e con la possibilità di accesso da punti remoti; la tecnologia esistente ha ormai messo a disposizione di chiunque, tramite terminali, personal e mini computer, modem, linee pubbliche e private fonia/dati, la possibilità di accedere a banche di dati centralizzate, con spesa relativamente limitata e risultati soddisfacenti.

Lo scambio di informazioni in una rete (trasmissione dati) è soggetto a diverse minacce potenziali:

- introduzione passiva
  (cioè rivelazione non autorizzata di informazioni)
- introduzione attiva
  (cioè modifica, estrazione o inserimento non autorizzato di informazioni).

L'unica protezione possibile contro l'intrusione passiva è la cifratura delle informazioni da trasmettere.

Non va dimenticato che sicurezza e protezione devono riguardare le informazioni non solo nella fase elaborativa e di archiviazione, ma anche e soprattutto nel momento del loro trasferimento all'esterno tramite tecniche di data communication.

È in questa fase che l'informazione è particolarmente vulnerabile e va quindi maggiormente protetta.

I controlli sono tanto più importanti in quanto molti dispositivi per la trasmissione sono di proprietà e sono quindi controllati da organizzazioni diverse da quelle che inviano e ricevono i dati.

I controlli sulla trasmissione regolano tre aree e cioè:

- la fase di introduzione (input) del messaggio
- la sua trasmissione in rete
- il suo ricevimento da parte del destinatario

come è evidenziato in fig. 4.

I controlli sulla trasmissione sono necessari per assicurare che l'integrità del dato non venga persa o compromessa dal momento della immissione al terminale fino alla sua ricezione sull'elaboratore.

Durante il tempo di trasmissione i dati sono normalmente fuori dal controllo dell'ente che li origina e dell'ente che li elabora (essendo le linee di trasmissione generalmente affittate); viene a mancare un controllo fisico, per cui l'uso adeguato dei controlli (parità longitudinale e trasversale ecc.) è importante per garantire l'accuratezza e la completezza del dato trasmesso attraverso il sistema di comunicazione.

Il compito dell'EDP auditor è anche quello di vigilare sull'uso corretto e sicuro di tali strumenti che compongono le reti trasmissione dati.

Al riguardo l'EDP auditor dovrà verificare che:

- esista un adeguato controllo degli accessi alla rete con registrazione dei tentativi indebiti
- esista una netta separazione delle funzioni tra gestione del software e controllo della rete
- agli utenti periferici sia consentito solo l'uso di comandi autorizzati
- agli utenti periferici con funzione di data collection (che provvedono all'invio di lotti di messaggi in un'unica soluzione mediante file transfer) siano controllati i singoli messaggi ai fini della autorizzazione e della sicurezza prima del loro utilizzo, al fine di prevenire il "cavallo di Troia"
- il dispositivo di crittografia dei messaggi sia di tipo a variazione dinamica contemplando il log di audit che permetta di rilevare i periodi di eventuali perdite di sincronismo con conseguente trasmissione in chiaro anzichè in cifrato.
- l'hardware remoto non disponga di compilatori e che i programmi forniti all'utente siano esclusivamente in formato eseguibile. Il sorgente deve essere riservato so-

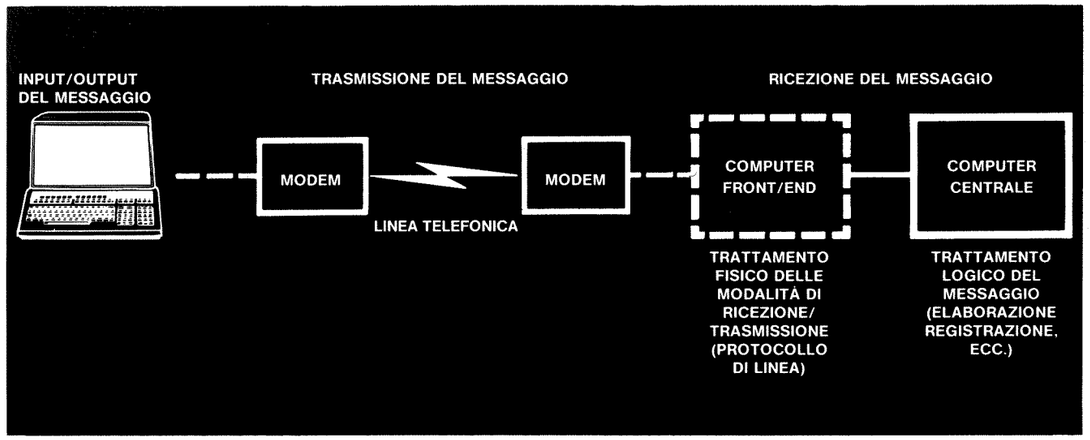


Fig. 4 - Aree di controlli nella trasmissione di informazioni.

lo ai responsabili dello sviluppo e manutenzione delle applicazioni

- gli operatori periferici siano adeguatamente addestrati e dispongano di sufficiente documentazione operativa
- il linguaggio operativo di rete NCL (network control language) utilizzato dall'utenza periferica abbia protetto con password il software di supervisione
- sia limitato il turnover degli utenti con disponibilità di sostituti altrettanto addestrati
- eventuali malfunzionamenti locali non pregiudichino il funzionamento della stazione remota o addirittura dell'intera rete e che vengano presi tempestivamente provvedimenti per risolvere l'anomalia
- esista un "Ispettorato" o Ente analogo locale in grado di controllare l'operato degli utenti periferici.

6.5. Controlli espletati dagli utenti

L'utente può definirsi l'utilizzatore finale delle risorse EDP (H/W, S/W). Egli collabora con l'EDP nella fase di stesura delle analisi di fattibilità e utilizza il sistema informativo al fine di ottenere i dati nella forma e nei tempi tali da consentire la migliore gestione delle informazioni da parte dell'azienda allo scopo di raggiungere gli obiettivi prefissati dalla direzione.

Per controlli degli utenti si intendono quei controlli manuali espletati sui dati sottoposti ad elaborazione che possono nascondere attacchi al patrimonio aziendale con l'eliminazione di operazioni attive o l'inserimento di operazioni passive.

Altri controlli consistono in procedure manuali svolte dagli utenti per verificare i risultati di elaborazioni automatiche.

Controlli di questo tipo sono ad esempio, quelli che verificano la completezza delle fatture di vendita stampate dal computer con un documento accompagnato dalle fatture stesse.

7. Conclusioni

È da tener presente che l'attività di molte aziende (nonchè l'economia di molte nazioni) è basata su sistemi computerizzati.

Scopo principale dell'"auditor EDP" è di garantire al management una adeguata sicurezza dell'area EDP, nel rispetto delle normative aziendali. A tale fine l'auditor EDP, agendo attraverso analisi e indagini, deve valutare le possibili situazioni di rischio e prendere gli opportuni provvedimenti per la sicurezza.

Ad ogni rischio deve contrapporsi l'adeguato provvedimento.

Affinchè la funzione EDP auditing possa efficaciemente ed efficientemente svolgere il proprio compito istituzionale di controllo per conto dell'"alta direzione", è necessario che essa sia collocata nell'ambito della struttura aziendale in posizioni che assicurino agli auditor l'autonomia e l'obiettività di giudizio sulle aree aziendali oggetto di controllo.

8. Bibliografia

[1] BUSSOLATI U., MARTELLA G., PETRONIO C., *La sicurezza dei sistemi informativi,* F. Angeli, Milano, 1981.

[2] CHAMBERS A., *Internal auditing: theory and practice,* Pitman, London, 1981.

[3] CIRTIN A., *Risk analysis of internal control procedures,* The internal auditor, June 1982.

[4] MARASCO R., *Il problema della sicurezza nei sistemi informatici,* Ufficio stile, Febbr. 1985.

[5] MARASCO R., *Il computer crime,* Consulenza informatica, II, 1, 1986.

[6] MARASCO R., *Innovazione tecnologica nel mondo bancario, trasferimento elettronico dei fondi, swift e sicurezza,* Consulenza informatica, II, 3, 1986.

[7] MARASCO R., *La sicurezza nella trasmissione dati. I quattro codici dell'auditing,* IPSOA Informatica, Milano, 1986.

[8] MARASCO R., *La sicurezza dei dati nell'EDP,* Risk management, 1983.

[9] MONALDO G., *Manuale di revisione,* IPSOA, Milano 1984.

[10] OCCHINI G., *L'informatica nella gestione aziendale,* F. Angeli, Milano, 1982.

[11] PARKER D.B., *Computer security differences for accidental and intentionally caused losses,* Stanford Research Institute, 1978.

Collana dei Quaderni di Informatica

La collana di testi "Quaderni di Informatica" rappresenta la proiezione sul piano librario della presente rivista. La direzione scientifica della collana (che è edita da F. Angeli) è del "Centro Studi Informatica e Automazione" della Honeywell I.S.I. A titolo di rassegna sintetica, riproduciamo qui di seguito le copertine dei volumi finora usciti, che coprono alcune delle più importanti aree dell'informatica.